

(11)特許出願公表番号
特表2003-507817
(P2003-507817A)

(43)公表日 平成15年2月25日(2003.2.25)

(51)Int.Cl. ⁷	識別記号	F I	7-73-7* (参考)
G 0 6 F 15/16	6 4 0	G 0 6 F 15/16	6 4 0 A 5 B 0 4 5
15/177	6 7 0	15/177	6 7 0 A
			6 7 0 B
			6 7 0 C
	6 7 4		6 7 4 B
		審査請求 未請求	予備審査請求 有 (全 74 頁)

(21)出願番号	特願2001-519281(P2001-519281)
(86)(22)出願日	平成12年8月17日(2000.8.17)
(85)翻訳文提出日	平成14年2月20日(2002.2.20)
(86)国際出願番号	PCT/US00/22783
(87)国際公開番号	WO01/014987
(87)国際公開日	平成13年3月1日(2001.3.1)
(31)優先権主張番号	60/150,394
(32)優先日	平成11年8月23日(1999.8.23)
(33)優先権主張国	米国(US)
(31)優先権主張番号	09/502,170
(32)優先日	平成12年2月11日(2000.2.11)
(33)優先権主張国	米国(US)

(71)出願人 テラスプリング・インコーポレーテッド
アメリカ合衆国 カリフォルニア州
94538 フレモント ミルモント ドライ
ブ 48800

(72)発明者 アシャー・アズイズ
アメリカ合衆国 カリフォルニア州
94555 フレモント タナガー コモン
4180

(72)発明者 トム・マークソン
アメリカ合衆国 カリフォルニア州
94402 サン マテオ マウンズ ロード
ナンバー206 30

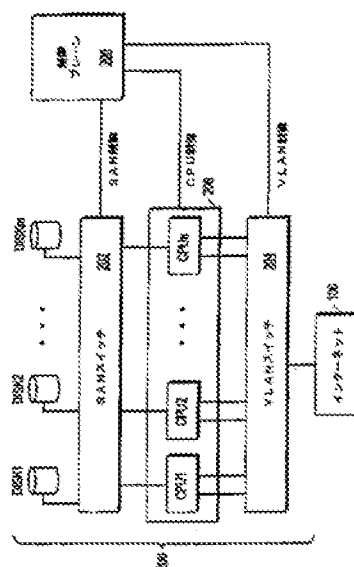
(74)代理人 弁理士 渡辺 喜平 (外1名)

● 教育行政

(54)【発明の名称】 拡張可能なコンピューティング・システム

(57) 〔要約〕

動的なサイズ変更が可能であり、高度にスケーラブルかつ使用可能なサーバ・ファームを提供する方法および装置について開示する。仮想サーバ・ファーム（VSF）は、いったん物理的に構築され、次いでオン・デマンドで様々な組織用のVSFに論理的に分割される、大規模コンピューティング構造（「コンピューティング・グリッド」）から作成される。各組織は、VSFの管理制御を独自に保持する。VSFはコンピューティング・グリッド内で動的にファイアウォールを設置する。VSF内での要素の割振りおよび制御は、特別な制御ポートを介して、コンピューティング・グリッド内にあるすべてのコンピューティング要素、ネットワーク要素、および記憶要素に接続された、制御プレーンによって実行される。各VSFの内部トポロジは、制御プレーンの制御下にある。単層のWebサーバまたは多層のWebサーバ、アプリケーション・サーバ、データベース・サーバなどの構成を含む多くの異なる構成で、VSFを構築するために、物理的な配線変更は必要ない。多層VSFの各層（たとえばWebサーバ層、アプリケーション・サ



【特許請求の範囲】

【請求項1】 プロセッサのセットの中から、該プロセッサのサブセットを選択するステップと、

前記プロセッサのサブセット中の各プロセッサ同士が論理的に結合するように、第1のスイッチング・システムを動作させる命令を生成するステップと、

記憶デバイスのセットの中から、該記憶デバイスのサブセットを選択するステップと、

前記記憶デバイスのサブセット中の各記憶デバイス同士が、互いに及び前記プロセッサのサブセットと、論理的に結合するように、第2のスイッチング・システムを動作させる命令を生成するステップと、
の各ステップを備えるデータ処理方法。

【請求項2】 前記プロセッサのセットの中から該プロセッサのサブセットを選択する前記ステップが、利用可能な中央処理装置のプールの中から中央処理装置のサブセットを選択するステップを備える請求項1に記載の方法。

【請求項3】 前記プロセッサのセットの中から該プロセッサのサブセットを選択する前記ステップが、利用可能な中央処理装置のプールの中から中央処理装置のサブセットを選択するステップを備え、

前記中央処理装置のそれぞれが、仮想ローカル・エリア・ネットワーク・スイッチからの命令を受信するよう構成された第1及び第2のネットワーク・インタフェースと、記憶エリア・ネットワーク・スイッチを介して前記記憶デバイスのサブセットに接続されるよう構成された記憶インタフェースとを含んでいる請求項1に記載の方法。

【請求項4】 前記プロセッサのサブセット中の各プロセッサ同士が論理的に結合するように、第1のスイッチング・システムを動作させる命令を生成する前記ステップが、仮想ローカル・エリア・ネットワーク・スイッチを前記プロセッサに結合して、前記仮想ローカル・エリア・ネットワーク・スイッチで前記サブセット内の前記プロセッサ同士が結合するようにさせる命令を生成するステップを備える請求項1に記載の方法。

【請求項5】 記憶デバイスのセットの中から該記憶デバイスのサブセット

を選択する前記ステップが、利用可能な記憶デバイスのプールの中から前記記憶デバイスのサブセットを選択するステップを備える請求項1に記載の方法。

【請求項6】 記憶デバイスのセットの中から該記憶デバイスのサブセットを選択する前記ステップが、利用可能な記憶デバイスのプールの中から前記記憶デバイスのサブセットを選択するステップを備え、

前記記憶デバイスのそれぞれが、仮想記憶エリア・ネットワーク・スイッチからの命令を受信するよう構成されたスイッチング・インタフェースを含んでいる請求項1に記載の方法。

【請求項7】 前記記憶デバイスのサブセット中の各記憶デバイス同士が論理的に結合するように、第2のスイッチング・システムを動作させる命令を生成する前記ステップが、仮想記憶エリア・ネットワーク・スイッチを前記記憶デバイスに結合して、前記仮想記憶エリア・ネットワーク・スイッチで前記サブセット内の前記記憶デバイス同士が結合するようにさせる命令を生成するステップを備える請求項1に記載の方法。

【請求項8】 前記プロセッサのセットの中から該プロセッサのサブセットを選択する前記ステップが、前記第1のスイッチング・システム、前記第2のスイッチング・システム及び前記プロセッサのセットに結合され、これらを制御するコントローラによって実行される請求項1に記載の方法。

【請求項9】 第1のデータ処理操作において使用する第1の仮想サーバ・ファームを生成するステップであって、該ステップが、前記プロセッサのセットの中から、前記第1のデータ処理操作を処理するための前記プロセッサの第1のサブセットを選択し、前記第1のスイッチング・システムで、第1の仮想ローカル・エリア・ネットワークにおける前記プロセッサの第1のサブセット内の前記プロセッサ同士が結合するようにさせる命令を生成し、前記記憶デバイスのセットの中から、前記第1のデータ処理の問題に係る情報を記憶するための記憶デバイスの第1のサブセットを選択し、更に、第2のスイッチング・システムで、前記記憶デバイスの第1のサブセット内の各記憶デバイス同士が、互いに及び第1の記憶エリア・ネットワーク・ゾーン内の前記プロセッサの第1のサブセットと、論理的に結合するようにさせる命令を生成することによって達成されるものと

、

第2のデータ処理操作において使用する第2の仮想サーバ・ファームを生成するステップであって、該ステップが、前記プロセッサのセットの中から、前記第2のデータ処理操作を処理するための前記プロセッサの第2のサブセットを選択し、前記第1のスイッチング・システムで、第2の仮想ローカル・エリア・ネットワークにおける前記プロセッサの第2のサブセット内の前記プロセッサ同士が結合するようにさせる命令を生成し、前記記憶デバイスのセットの中から、前記第2のデータ処理の問題に係る情報を記憶するための記憶デバイスの第2のサブセットを選択し、更に、第2のスイッチング・システムで、前記記憶デバイスの第2のサブセット内の各記憶デバイス同士が、互いに及び第2の記憶エリア・ネットワーク・ゾーン内の前記プロセッサの第2のサブセットと、論理的に結合するようにさせる命令を生成することによって達成されるものと、
を更に備え、前記命令が、前記プロセッサの第1のサブセットを、前記プロセッサの第2のサブセット及び前記記憶デバイスの第2のサブセットから保守上分離するものである請求項1に記載の方法。

【請求項10】 前記プロセッサのセットの中から、追加のプロセッサを選択するステップと、

前記第1のスイッチング・システムが、前記追加のプロセッサを、前記プロセッサのサブセット内のプロセッサに論理的に結合するように動作する命令を生成するステップと、
を更に備える請求項1に記載の方法。

【請求項11】 前記プロセッサのサブセットの中から、該サブセットから除去する特定のプロセッサを選択するステップと、

前記第1のスイッチング・システムが、前記プロセッサのサブセットから前記特定のプロセッサを論理的に分離するように動作する命令を生成するステップと、
、
を更に備える請求項1に記載の方法。

【請求項12】 前記プロセッサのサブセットの中から、該サブセットから除去する特定のプロセッサを選択するステップと、

前記第1のスイッチング・システムが、前記プロセッサのサブセットから前記特定のプロセッサを論理的に分離するように動作する命令を生成するステップと、

前記利用可能なプロセッサのプール内に、前記特定のプロセッサを論理的に配置するステップと、
を更に備える請求項2に記載の方法。

【請求項13】 前記プロセッサの第1のサブセットの中から、該第1のサブセットから除去する特定のプロセッサを選択するステップと、

前記第1のスイッチング・システムが、前記プロセッサの第1のサブセットから前記特定のプロセッサを論理的に分離するように動作する命令を生成するステップと、

前記第1のスイッチング・システムが、前記プロセッサの第2のサブセットへ前記特定のプロセッサを論理的に追加するように動作する命令を生成するステップと、
を更に備える請求項9に記載の方法。

【請求項14】 前記プロセッサの全てを利用可能なプロセッサのアイドル・プールへ最初に割り当てるステップを更に備える請求項1に記載の方法。

【請求項15】 前記プロセッサのサブセットにより経験されたリアルタイム負荷に応じて、1又は複数のプロセッサを、前記プロセッサのサブセットへ又は該サブセットから、動的に論理的に追加し又は除去するステップを更に備える請求項1に記載の方法。

【請求項16】 前記記憶デバイスのサブセットにより経験されたリアルタイム負荷に応じて、1又は複数の記憶デバイスを、前記記憶デバイスのサブセットへ又は該サブセットから、動的に論理的に追加し又は除去するステップを更に備える請求項1に記載の方法。

【請求項17】 前記第1のスイッチング・システムのインタフェースを外部ネットワークへ結合するステップを更に備え、これにより前記プロセッサのサブセットが前記外部ネットワークからの要求に対して応答するものとなる請求項1に記載の方法。

【請求項18】 前記プロセッサのサブセットにより経験されたリアルタイム負荷に応じて、前記プロセッサのサブセットへ追加のプロセッサを論理的に追加するステップと、

前記追加のプロセッサを再起動するステップと、

前記追加のプロセッサ内に、前記記憶デバイスのサブセット内の1つの記憶デバイスの予め定義された場所からオペレーティング・システム・イメージをロードするステップと、

の各ステップを更に備える請求項15に記載の方法。

【請求項19】 ロード・バランサを前記プロセッサのサブセット内のプロセッサに論理的に結合するステップと、

前記ロード・バランサに、前記プロセッサのサブセット内のプロセッサにより実行されるロード・バランサ処理を命令するステップと、

を更に備える請求項1に記載の方法。

【請求項20】 前記プロセッサのサブセットにより経験されたリアルタイム負荷に応じて、前記プロセッサのサブセットへ追加のプロセッサを動的に論理的に追加するステップを更に備え、該ステップが、

前記仮想ローカル・エリア・ネットワークへ、前記追加のプロセッサに関連付けられた前記第1のスイッチングシステムのそれらのインタフェース・ポートを追加するステップと、

前記追加のプロセッサの記憶エリア・ポートを、前記記憶エリア・ネットワーク・ゾーンへ追加するステップと、

により達成される請求項9に記載の方法。

【請求項21】 前記プロセッサのサブセットにより経験されたリアルタイム負荷に応じて、前記プロセッサのサブセットから1つのプロセッサを動的に論理的に除去するステップを更に備え、該ステップが、

前記仮想ローカル・エリア・ネットワークから、前記追加のプロセッサに関連付けられた前記第1のスイッチングシステムのそれらのインタフェース・ポートを除去するステップと、

前記1つのプロセッサの記憶エリア・ポートを、前記記憶エリア・ネットワー

ク・ゾーンから除去するステップと、
により達成される請求項9に記載の方法。

【請求項22】 前記第2のスイッチング・システムの1又は複数のポートを、前記コントローラによって使用されるプライベートな記憶エリア・ネットワーク・ゾーンに論理的に割り当てるステップであって、前記ポートが、前記コントローラに排他的に割り当てられた1つの前記記憶デバイスに関連付けられているものを更に備えた請求項8に記載の方法。

【請求項23】 前記プロセッサのサブセット内の各プロセッサを、予め定義され記憶された設計図と関連付けするステップであって、前記設計図が、複数の処理の役割の1つを、前記記憶デバイスのサブセット内の1つの記憶デバイスのブート・イメージと関連付けているものと、

前記プロセッサのサブセット内のプロセッサのそれぞれが、そのプロセッサの前記処理の役割に関連付けられた前記記憶デバイスからの前記ブート・イメージをロードし、実行するように動作する命令を生成するステップと、
を更に備える請求項1に記載の方法。

【請求項24】 第1のデータ処理操作において使用する第1の仮想サーバ・ファームを生成するステップであって、該ステップが、前記プロセッサのセットの中から、前記第1のデータ処理操作を処理するための前記プロセッサの第1のサブセットを選択し、前記第1のスイッチング・システムで、第1の仮想ローカル・エリア・ネットワークにおける前記プロセッサの第1のサブセット内の前記プロセッサ同士が結合するようにさせる命令を生成し、前記記憶デバイスのセットの中から、前記第1のデータ処理の問題に係る情報を記憶するための記憶デバイスの第1のサブセットを選択し、更に、第2のスイッチング・システムで、前記記憶デバイスの第1のサブセット内の各記憶デバイス同士が、互いに及び第1の記憶エリア・ネットワーク・ゾーン内の前記プロセッサの第1のサブセットと、論理的に結合するようにさせる命令を生成することによって達成されるものと、

第2のデータ処理操作において使用する第2の仮想サーバ・ファームを生成するステップであって、該ステップが、前記プロセッサのセットの中から、前記第

2のデータ処理操作を処理するための前記プロセッサの第2のサブセットを選択し、前記第1のスイッチング・システムで、第2の仮想ローカル・エリア・ネットワークにおける前記プロセッサの第2のサブセット内の前記プロセッサ同士が結合するようにさせる命令を生成し、前記記憶デバイスのセットの中から、前記第2のデータ処理の問題に係る情報を記憶するための記憶デバイスの第2のサブセットを選択し、更に、第2のスイッチング・システムで、前記記憶デバイスの第2のサブセット内の各記憶デバイス同士が、互いに及び第2の記憶エリア・ネットワーク・ゾーン内の前記プロセッサの第2のサブセットと、論理的に結合するようにさせる命令を生成することによって達成されるものと、
を更に備え、前記命令が、前記プロセッサの第1のサブセットを、前記プロセッサの第2のサブセット及び前記記憶デバイスの第2のサブセットから保守上分離するものであるデータ処理方法。

【請求項25】 複数のプロセッサと、
前記複数のプロセッサに接続された第1のスイッチング・システムと、
複数の記憶デバイスと、
前記複数の記憶デバイスに接続された第2のスイッチング・システムと、
前記第1のスイッチング・システム及び前記第2のスイッチング・システムに接続されたコントローラと、
前記複数のプロセッサの中から、該プロセッサのサブセットを選択する前記コントローラ内の手段と、
前記プロセッサのサブセット中の各プロセッサ同士が論理的に結合するように、第1のスイッチング・システムを動作させる命令を生成する前記コントローラ内の手段と、
前記複数の記憶デバイスの中から、該記憶デバイスのサブセットを選択する前記コントローラ内の手段と、
前記記憶デバイスのサブセット中の各記憶デバイス同士が、互いに及び前記プロセッサのサブセットと、論理的に結合するように、第2のスイッチング・システムを動作させる命令を生成する前記コントローラ内の手段と、
を備えるデータ処理システム。

【請求項26】 前記コントローラが、利用可能な中央処理装置のプールの中から中央処理装置のサブセットを選択することにより、前記プロセッサのセットの中から該プロセッサのサブセットを選択する手段を更に備える請求項25に記載のデータ処理システム。

【請求項27】 前記プロセッサのサブセットを選択する前記手段が、利用可能な中央処理装置のプールの中から中央処理装置のサブセットを選択する手段を備え、

前記中央処理装置のそれぞれが、仮想ローカル・エリア・ネットワーク・スイッチからの命令を受信するよう構成された第1及び第2のネットワーク・インタフェースと、記憶エリア・ネットワーク・スイッチを介して前記記憶デバイスのサブセットに接続されるよう構成された記憶インタフェースとを含んでいる請求項25に記載のデータ処理システム。

【請求項28】 前記プロセッサのサブセット中の各プロセッサ同士が論理的に結合するように、第1のスイッチング・システムを動作させる命令を生成する前記手段が、仮想ローカル・エリア・ネットワーク・スイッチを前記プロセッサに結合して、前記仮想ローカル・エリア・ネットワーク・スイッチで前記サブセット内の前記プロセッサ同士が結合するようにさせる命令を生成する手段を備える請求項25に記載のデータ処理システム。

【請求項29】 記憶デバイスのサブセットを選択する前記手段が、利用可能な記憶デバイスのプールの中から前記記憶デバイスのサブセットを選択する手段を備える請求項25に記載のデータ処理システム。

【請求項30】 記憶デバイスのサブセットを選択する前記手段が、利用可能な記憶デバイスのプールの中から前記記憶デバイスのサブセットを選択する手段を備え、

前記記憶デバイスのそれぞれが、仮想記憶エリア・ネットワーク・スイッチからの命令を受信するよう構成されたスイッチング・インタフェースを含んでいる請求項25に記載のデータ処理システム。

【請求項31】 前記各記憶デバイス同士が論理的に結合するように、第2のスイッチング・システムを動作させる命令を生成する前記手段が、仮想記憶エ

リア・ネットワーク・スイッチを前記記憶デバイスに結合して、前記仮想記憶エリア・ネットワーク・スイッチで前記サブセット内の前記記憶デバイス同士が結合するようにさせる命令を生成する手段を備える請求項25に記載のデータ処理システム。

【請求項32】 第1のデータ処理操作において使用する第1の仮想サーバ・ファームであって、前記プロセッサのセットの中から、前記第1のデータ処理操作を処理するための前記プロセッサの第1のサブセットを選択し、前記第1のスイッチング・システムで、第1の仮想ローカル・エリア・ネットワークにおける前記プロセッサの第1のサブセット内の前記プロセッサ同士が結合するようにさせる命令を生成し、前記記憶デバイスのセットの中から、前記第1のデータ処理の問題に係る情報を記憶するための記憶デバイスの第1のサブセットを選択し、更に、第2のスイッチング・システムで、前記記憶デバイスの第1のサブセット内の各記憶デバイス同士が、互いに及び第1の記憶エリア・ネットワーク・ゾーン内の前記プロセッサの第1のサブセットと、論理的に結合するようにさせる命令を生成することによって生成されるものと、

第2のデータ処理操作において使用する第2の仮想サーバ・ファームであって、前記プロセッサのセットの中から、前記第2のデータ処理操作を処理するための前記プロセッサの第2のサブセットを選択し、前記第1のスイッチング・システムで、第2の仮想ローカル・エリア・ネットワークにおける前記プロセッサの第2のサブセット内の前記プロセッサ同士が結合するようにさせる命令を生成し、前記記憶デバイスのセットの中から、前記第2のデータ処理の問題に係る情報を記憶するための記憶デバイスの第2のサブセットを選択し、更に、第2のスイッチング・システムで、前記記憶デバイスの第2のサブセット内の各記憶デバイス同士が、互いに及び第2の記憶エリア・ネットワーク・ゾーン内の前記プロセッサの第2のサブセットと、論理的に結合するようにさせる命令を生成することによって生成されるものと、

を更に備え、前記命令が、前記プロセッサの第1のサブセットを、前記プロセッサの第2のサブセット及び前記記憶デバイスの第2のサブセットから保守上分離するものである請求項25に記載のデータ処理システム。

【請求項33】 前記プロセッサのセットの中から、追加のプロセッサを選択する手段と、

前記第1のスイッチング・システムが、前記追加のプロセッサを、前記プロセッサのサブセット内のプロセッサに論理的に結合するように動作する命令を生成する手段と、

を更に備える請求項25に記載のデータ処理システム。

【請求項34】 前記プロセッサのサブセットの中から、該サブセットから除去する特定のプロセッサを選択する手段と、

前記第1のスイッチング・システムが、前記プロセッサのサブセットから前記特定のプロセッサを論理的に分離するように動作する命令を生成する手段と、

を更に備える請求項25に記載のデータ処理システム。

【請求項35】 前記プロセッサのサブセットの中から、該サブセットから除去する特定のプロセッサを選択する手段と、

前記第1のスイッチング・システムが、前記プロセッサのサブセットから前記特定のプロセッサを論理的に分離するように動作する命令を生成する手段と、

前記利用可能なプロセッサのプール内に、前記特定のプロセッサを論理的に配置する手段と、

を更に備える請求項26に記載のデータ処理システム。

【請求項36】 前記プロセッサの第1のサブセットの中から、該第1のサブセットから除去する特定のプロセッサを選択する手段と、

前記第1のスイッチング・システムが、前記プロセッサの第1のサブセットから前記特定のプロセッサを論理的に分離するように動作する命令を生成する手段と、

前記第1のスイッチング・システムが、前記プロセッサの第2のサブセットへ前記特定のプロセッサを論理的に追加するように動作する命令を生成する手段と、

、

を更に備える請求項28に記載のデータ処理システム。

【請求項37】 前記プロセッサの全てを利用可能なプロセッサのアイドル・プールへ最初に割り当てる手段を更に備える請求項25に記載のデータ処理シ

ステム。

【請求項38】 前記プロセッサのサブセットにより経験されたリアルタイム負荷に応じて、1又は複数のプロセッサを、前記プロセッサのサブセットへ又は該サブセットから、動的に論理的に追加し又は除去する手段を更に備える請求項25に記載のデータ処理システム。

【請求項39】 前記記憶デバイスのサブセットにより経験されたリアルタイム負荷に応じて、1又は複数の記憶デバイスを、前記記憶デバイスのサブセットへ又は該サブセットから、動的に論理的に追加し又は除去する手段を更に備える請求項25に記載のデータ処理システム。

【請求項40】 前記コントローラが、前記サブセット内の各プロセッサの処理負荷を表す情報を生成する負荷監視装置を更に備える請求項25に記載のデータ処理システム。

【請求項41】 前記第1のスイッチング・システムが、偽造不可能なポート識別子を有する仮想ローカル・エリア・ネットワーク・スイッチを備える請求項25に記載のデータ処理システム。

【請求項42】 前記第2のスイッチング・システムが、記憶エリア・ネットワーク・スイッチを備え、前記記憶デバイスのサブセットが、記憶エリア・ネットワーク・ゾーン内で論理的に組織化され、更に、前記記憶エリア・ネットワーク・スイッチが、前記記憶デバイスのサブセットへのアクセスを、前記プロセッサのサブセット内のそれらのプロセッサに対してのみ許可するものである請求項25に記載のデータ処理システム。

【請求項43】 前記第2のスイッチング・システムが、記憶エリア・ネットワーク・スイッチを備え、前記記憶デバイスのサブセットが、記憶エリア・ネットワーク・ゾーン内で論理的に組織化され、更に、前記記憶エリア・ネットワーク・スイッチが、前記記憶デバイスのサブセットへのアクセスを、1又は複数のファイバ・チャネル・スイッチを用いて、前記プロセッサのサブセット内のそれらのプロセッサに対してのみ許可するものである請求項25に記載のデータ処理システム。

【請求項44】 前記第2のスイッチング・システムが、記憶エリア・ネッ

トワーク・スイッチを備え、前記記憶デバイスのサブセットが、記憶エリア・ネットワーク・ゾーン内で論理的に組織化され、更に、前記プロセッサのサブセットが、ロード・バランサ又はファイアウォールを介して外部ネットワークへ結合されている請求項25に記載のデータ処理システム。

【請求項45】 前記第1のスイッチング・システムの制御ポートを含み、且つ前記第2のスイッチング・システムの制御ポートを含む、サブネットワーク内で論理的に相互接続された複数のコントローラを更に備える請求項25に記載のデータ処理システム。

【請求項46】 その時のプロセッサ負荷、ネットワーク負荷又は記憶負荷を表すリアルタイム情報を受信するために、前記プロセッサのサブセット内の各プロセッサを周期的にポーリングするよう構成されると共に、前記コントローラのそれぞれに前記情報を伝達するよう構成されたエージェント・コントローラを更に備える請求項45に記載のデータ処理システム。

【請求項47】 前記第2のスイッチング・システムの1又は複数のポートを、前記コントローラによって使用されるプライベートな記憶エリア・ネットワーク・ゾーンに論理的に割り当てるステップであって、前記ポートが、前記コントローラに排他的に割り当てられた1つの前記記憶デバイスに関連付けられているものを更に備えた請求項45に記載のデータ処理システム。

【請求項48】 予め定義され記憶された複数の設計図であって、該各設計図が、複数の処理の役割の1つを、前記記憶デバイスのサブセット内の1つの記憶デバイスのブート・イメージと関連付けているものと、

前記プロセッサのサブセット内のプロセッサのそれぞれを、前記設計図の1つと関連付け、前記プロセッサのサブセット内のプロセッサのそれぞれが、そのプロセッサの前記処理の役割に関連付けられた前記記憶デバイスからの前記ブート・イメージをロードし、実行するように動作させる手段と、
を更に備える請求項25に記載のデータ処理システム。

【請求項49】 プロセッサの複数のサブセット内で論理的に組織化された複数のプロセッサであって、前記サブセットのそれぞれが、複数の仮想ローカル・エリア・ネットワークの1つとして論理的に組織化されているものと、

前記サブセットの1つによって使用されるデータ及び命令を記憶するために、前記複数の仮想ローカル・エリア・ネットワークに結合され、記憶エリア・ネットワーク内で論理的に組織化された複数の記憶デバイスと、

前記複数の仮想ローカル・エリア・ネットワーク及び前記記憶エリア・ネットワークに結合された制御プレーンであって、リアルタイムで発生する処理負荷状況の変化及び記憶需要の変化に応じて、前記サブセットに対しプロセッサを動的に追加し又は除去すると共に、前記記憶エリア・ネットワークに対し記憶デバイスを動的に追加し又は除去するよう構成されたものと、
を備える仮想コンピューティング・システム。

【請求項50】 予め定義され記憶された複数の設計図であって、該各設計図が、複数の処理の役割の1つを、前記記憶デバイスのサブセット内の1つの記憶デバイスのブート・イメージと関連付けているものと、

前記プロセッサのサブセット内のプロセッサのそれぞれを、前記設計図の1つと関連付け、前記プロセッサのサブセット内のプロセッサのそれぞれが、そのプロセッサの前記処理の役割に関連付けられた前記記憶デバイスからの前記ブート・イメージをロードし、実行するように動作させる手段と、
を更に備える請求項49に記載の仮想コンピューティング・システム。

【請求項51】 拡張可能なコンピューティング・システムで使用するデータ処理のための1又は複数の命令のシーケンスを格納したコンピュータ読み取り可能な媒体であって、1又は複数のプロセッサによる前記1又は複数の命令のシーケンスの実行が、該1又は複数のプロセッサに、

プロセッサのセットの中から、該プロセッサのサブセットを選択するステップと、

前記プロセッサのサブセット中の各プロセッサ同士が論理的に結合するように、第1のスイッチング・システムを動作させる命令を生成するステップと、

記憶デバイスのセットの中から、該記憶デバイスのサブセットを選択するステップと、

前記記憶デバイスのサブセット中の各記憶デバイス同士が、互いに及び前記プロセッサのサブセットと、論理的に結合するように、第2のスイッチング・シス

テムを動作させる命令を生成するステップと、
の各ステップの動作を生じさせるコンピュータ読み取り可能な媒体。

【発明の詳細な説明】**【0001】**

【発明の技術分野】本発明は、一般にデータ処理に関する。より詳細には、本発明は、拡張可能、フレキシブル、かつスケーラブルなコンピューティング・システムを提供するための方法、装置、およびメカニズムに関する。

【0002】

【発明の背景】今日のWebサイトおよび他のコンピュータ・システムの構築者は、多くのシステム・プランニングに関する問題进行处理しなければならない。こうした問題には、通常の成長、予測される、または予測できないピーク時需要、サイトの使用可能性およびセキュリティなどを見込んだ処理能力プランニングが含まれる。Web上にサービスを提供しようとする企業は新しいビジネスおよびサービス・モデルを携え、その分野を革新し、首位に立とうと考えるが、それを実現させるためには大規模なWebサイトを設計、構築、および運営することの少なからぬ複雑さに対処しなければならない。これには、サイトを動作可能な状態のままで成長させ、かつ拡張する必要がある。

【0003】これをすべて実行するためには、潜在的に大規模かつ複雑である可能性のあるこうしたサイトのエンジニアリングおよび運営のできる技術者を探し雇い入れることが必要である。こうした大規模サイトの設計、構築、および運営は、多くの組織にとって中核業務ではまったくないため、これが難題を生み出している。

【0004】こうした問題の対応策の1つが、他企業の他のWebサイトと共用で、第三者のサイトで企業Webサイトをホストすることである。こうしたアウトソーシング施設は、現在、Exodus (TM)、AboveNet (TM)、GlobalCenter (TM)などの企業から利用することができる。これらの施設は、物理的なスペース、および冗長ネットワークや電力施設を提供しており、その結果、企業の顧客またはユーザはこれらを提供する必要がない。ネットワークおよび電力施設は、多くの企業または顧客が共用する。

【0005】しかし、これら施設のユーザには、自らの施設の構築、運営、および展開を行う中で、自らのコンピューティング・インフラストラクチャに

関する多くの作業を行なう必要が依然として残されている。こうした施設でホストされる企業の情報技術管理者は、施設で自分のコンピューティング機器の選択、導入、構成、および保守を行う責任を負う。管理者は、資源プランニングおよびピーク時処理能力の扱いなどの困難な問題に、なおも直面しなければならない。

【0006】たとえばアウトソーシング企業がコンピューティング施設も提供している場合であっても（たとえばD i g e x (TM)）、成長させるには同じく手操作の、誤りを犯しやすい管理ステップが含まれているので、アウトソーシング企業にとって拡張および成長は決して容易ではない。さらに、予測できないピーク時の需要に対する処理能力プランニングという問題も残されたままである。

【0007】さらに、Webサイトには、それぞれ異なる要件がある。たとえば、特定のWebサイトが、個別に管理および制御する機能を必要とする場合がある。サービス・プロバイダを共用している他のすべてのサイトからWebサイトを隔離する、特定のタイプまたはレベルのセキュリティを必要とするものもある。ほかの場所にある企業イントラネットとの安全な接続を必要とするものもある。

【0008】また、内部トポロジの点でも様々に異なるWebサイトがある。サイトの中には、単にWebのロード・バランサによってロード・バランシングされた一並びのWebサーバを含むだけのものがある。好適なロード・バランサは、C i s c o S y s t e m s , I n c . (TM)のL o c a l D i r e c t o r (TM)、F 5 L a b s (TM)のB i g I P (TM)、A l t e o n (TM)のW e b D i r e c t o r (TM)などである。他のサイトには、多層様式で構築されるものもあり、それによって一並びのWebサーバがH T T P (H y p e r t e x t T r a n s f e r P r o t o c o l) 要求を処理するが、大量のアプリケーション論理は別のアプリケーション・サーバで実施される。そして今度はこれらのアプリケーション・サーバを、元のデータベース・サーバの層に接続する必要がある。

【0009】これらの異なる構成シナリオの一部が、図1A、図1B、お

よび図1Cに示されている。図1Aは、CPU102およびディスク104を含む単一のマシン100を含む、単純なWebサイトを示す構成図である。マシン100は、インターネット106として知られるグローバルなパケット交換データ・ネットワークまたは他のネットワークに結合される。マシン100は、前述の種類の共用設備に収容することができる。

【0010】図1Bは、複数のWebサーバWSA、WSB、WSCを含む1層のWebサーバ・ファーム110を示す構成図である。それぞれのWebサーバが、インターネット106に結合されたロード・バランサ112に結合される。ロード・バランサは、各サーバにかかる処理負荷の均衡を保つためにサーバ間のトラフィックを分割する。ロード・バランサ112は、Webサーバを無許可のトラフィックから守るためのファイアウォールを含むかまたはこれと結合することもできる。

【0011】図1Cは、WebサーバW1、W2などの層、アプリケーション・サーバA1、A2などの層、およびデータベース・サーバD1、D2などの層を含む、3層のサーバ・ファーム120を示す。Webサーバは、HTTP要求を処理するために提供される。アプリケーション・サーバは、大半のアプリケーション論理を実行する。データベース・サーバは、データベース管理システム(DBMS)ソフトウェアを実行する。

【0012】構築する必要がある種類のWebサイトのトポロジの多様性を考えれば、大規模なWebサイトを構築するには、それぞれを特注構築することが唯一の方法であるのは明らかであろう。実際のところ、これが従来方法であった。多くの組織が同じ問題について個別に苦労を重ね、それぞれのWebサイトをゼロから特注構築してきた。これは非効率的であり、様々な企業で膨大な量の作業を重複して行うことになる。

【0013】従来方法に付随するもう1つの問題は、資源および処理能力のプランニングである。Webサイトは、日によって、あるいは1日の中でも時間によって、かなり異なるレベルのトラフィックを受信する可能性がある。トラフィックのピーク時には、Webサイトのハードウェアまたはソフトウェアは、オーバロードが原因で妥当な時間内に要求に応答できない場合がある。その他

の時間には、Webサイトのハードウェアまたはソフトウェアは、処理能力過剰で十分に活用できない可能性がある。従来の方法では、過度の経費をかけずにあるいは処理能力過剰とならないように、ピーク時のトラフィックを処理できるだけのハードウェアとソフトウェアの均衡を見出すことは、難しい問題である。多くのWebサイトでは、適切な均衡を見出しておらず、絶えず処理能力不足または処理能力過剰となっている。

【0014】他の問題は、人的な誤りによって引き起こされる障害である。手作業で構築されたサーバ・ファームを使用する現在の方法に存在する大きな潜在的危険は、新しいサーバを作動中のサーバ・ファームに組み入れる際に発生した人的誤りが、サーバ・ファームの誤動作を引き起こし、その結果、そのWebサイトのユーザにサービスを提供できなくなる可能性があるということである。

【0015】前述の内容に基づき、この分野では、特別注文による構築を必要とせず、オン・デマンドで即時にかつ容易に拡張できるコンピューティング・システムを提供するように改善された方法および装置が明らかに必要である。

【0016】さらに、トラフィック・スループットの変化を反映して、必要に応じて展開したり隠したりすることのできる、複数の分離された処理ノードの作成をサポートするコンピューティング・システムも必要である。他の必要性は、本書に示した開示で明らかになる。

【0017】

【発明の概要】前述の必要性および目的、ならびに以下の記述から明らかになるであろう他の必要性および目的は、本発明によって達成されるものであって、本発明の一態様には、大規模コンピューティング構造（「コンピューティング・グリッド」）に基づいて、高度にスケーラブルであり、高度に使用可能および確実なデータ処理サイトを作成するための方法および装置が含まれる。コンピューティング・グリッドは、いったん物理的に構築され、次いでオン・デマンドで様々な組織に論理的に分割される。コンピューティング・グリッドには、1つまたは複数のVLANスイッチおよび1つまたは複数の記憶域ネットワーク（SAN）スイッチに結合された、多数のコンピューティング要素が含まれる。複数の

記憶デバイスがS A Nスイッチに結合され、適切なスイッチング論理およびコマンドを介して、1つまたは複数のコンピューティング要素に選択的に結合することができる。V L A Nスイッチの1ポートが、インターネットなどの外部ネットワークに結合される。監視メカニズム、レイヤ、マシン、またはプロセスが、V L A NスイッチおよびS A Nスイッチに結合される。

【0018】第一に、すべての記憶デバイスおよびコンピューティング要素がアイドル・プールに割り当てられる。プログラムの制御の下で、監視メカニズムが、V L A NスイッチおよびS A Nスイッチのポートを1つまたは複数のコンピューティング要素および記憶デバイスに結合するために、これらを動的に構成する。その結果、こうした要素およびデバイスはアイドル・プールから論理的に除去され、1つまたは複数の仮想サーバ・ファーム(V S F)の一部となる。それぞれのV S Fコンピューティング要素が、コンピューティング要素がブートストラップの動作および作成を実行するために使用可能なブート・イメージを含む記憶デバイスに向けられるか、またはそうでなければ関連付けられる。

【0019】コンピューティング・グリッドをいったん物理的に構築し、コンピューティング・グリッドの一部を様々な組織にオン・デマンドで確実にかつ動的に割り振ることで、各サイトを特別注文で構築する場合には達成が困難な、スケール・メリットが達成される。

【0020】本発明は、添付の図面において限定的なものでなく例示的なものとして示されており、この図面では、同じ要素が同じ参照番号で表されている。

【0021】

【好適な実施形態の説明】

拡張可能コンピューティング・システムを提供するための方法および装置について説明する。以下の説明では、本発明について完全に理解してもらうために、例示の目的で多数の特有の詳細について記載する。ただし、当分野の技術者には、本発明がこれらの特有の詳細なしに実施できることが明らかになる。他の場合には、本発明を必要以上に不明瞭にすることがないように、構成図にはよく知られた構造およびデバイスが示される。

【0022】

(仮想サーバ・ファーム (V S F))

一実施形態によれば、大規模コンピューティング構造（「コンピューティング・グリッド」）が提供される。コンピューティング・グリッドは、いったん物理的に構築し、次いでオン・デマンドで様々な組織に論理的に分割することができる。複数の企業または組織のそれぞれに、コンピューティング・グリッドの一部が割り振られる。各組織のコンピューティング・グリッドの論理部分が、仮想サーバ・ファーム (V S F) と呼ばれる。各組織は、そのV S Fの独立した管理制御を保持する。各V S Fは、サーバ・ファーム上に配置されたリアルタイム要求または他の要素に基づいて、C P Uの数、記憶容量およびディスク、ならびにネットワーク帯域幅によって動的に変更することができる。各V S Fは、たとえばすべて物理的に同じコンピューティング・グリッドから論理的に作成された場合であっても、あらゆる他の組織のV S Fから守られている。V S Fは、私設専用回線または仮想私設ネットワーク (V P N) のいずれかを使用して、イントラネットを他の組織のV S Fにさらすことなく、イントラネットに再接続することができる。

【0023】 組織は、たとえばコンピュータへの管理アクセスをフルに実施することができる（たとえばスーパー・ユーザまたはルート）か、またはコンピュータが接続されたローカル・エリア・ネットワーク (L A N) 上のすべてのトラフィックを監視することができる場合であっても、コンピューティング・グリッドの割り振られた部分、すなわちそのV S Fにあるデータおよびこれらのコンピューティング要素にのみアクセスすることが可能である。これは、V S Fのセキュリティ外辺部が動的に伸縮する、動的なファイアウォール・スキームを使用して達成される。

【0024】 各V S Fは、インターネット、イントラネット、またはエクストラネットを介してアクセス可能な組織のコンテンツおよびアプリケーションをホストするのに使用することができる。

【0025】 コンピューティング要素ならびにそれに関連付けられたネットワークワーキングおよび記憶要素の構成および制御は、コンピューティング・グリッド

ド内のいかなるコンピューティング要素を介しても直接アクセスすることのできない、監視メカニズムによって実施される。便宜上、本書では監視メカニズムを制御プレーンと呼び、1つまたは複数のプロセッサまたはプロセッサのネットワークを含むことができる。監視メカニズムは、監視プログラム、制御装置などを含むことができる。本明細書に記載するように、他の方法を使用することもできる。

【0026】制御プレーンは、ネットワーク内でまたは他の手段によって相互接続することができる1つまたは複数のサーバなどの、監視目的で割り当てられた完全に独立したコンピューティング要素セット上で動作する。これは、グリッド内のネットワークおよび記憶要素の特別な制御ポートまたはインターフェースを介して、コンピューティング・グリッドのコンピューティング、ネットワーク、および記憶要素上で制御動作を実行するものである。制御プレーンは、システムのスイッチング要素に物理インターフェースを提供し、システム内のコンピューティング要素の負荷を監視し、グラフィカル・ユーザ・インターフェースまたは他の好適なユーザ・インターフェースを使用する管理および監督機能を提供するものである。

【0027】制御プレーンを実行中のコンピュータは、コンピューティング・グリッド内にある（したがって任意の特有のVSF内にある）コンピュータには論理的に不可視であり、コンピューティング・グリッド内にあるかまたは外部コンピュータからの要素を介して、いかなる方法でも攻撃または破壊されることはない。制御プレーンのみが、コンピューティング・グリッド内のデバイス上にある制御ポートへの物理的接続を有し、特定のVSFでのメンバシップを制御する。コンピューティング内のデバイスは、これらの特殊な制御ポートを介してのみ構成することが可能であるため、コンピューティング・グリッド内のコンピューティング要素は、セキュリティ外辺部を変更するか、あるいは許可されていない記憶デバイスまたはコンピューティング・デバイスにアクセスすることができない。

【0028】したがってVSFは、組織が、大規模共用コンピューティング・インフラストラクチャ、すなわちコンピューティング・グリッドから動的に

作成された専用サーバ・ファームを含むように見える、コンピューティング施設を取り扱えるようにするものである。本明細書に記載のコンピューティング・アーキテクチャで結合された制御プレーンは、コンピューティング・グリッドのデバイスのハードウェアで実施されるアクセス制御メカニズムを介してプライバシーおよび安全性が保護されている専用サーバ・ファームを提供する。

【0029】各VSFの内部トポロジは、制御プレーンによって制御される。制御プレーンは、本明細書に記載のコンピュータ、ネットワーク・スイッチ、および記憶ネットワーク・スイッチの基本的な相互接続を採用し、これらを使用して様々なサーバ・ファーム構成を作成することができる。これらには、ロード・バランサがフロントエンドとなる単一層のWebサーバ・ファーム、ならびにWebサーバがアプリケーション・サーバに伝え、次にこれがデータベース・サーバに伝える多層構成が含まれるが、これらに限定されるものではない。様々なロード・バランシング、多層、およびファイアウォールの構成が可能である。

【0030】

(コンピューティング・グリッド)

コンピューティング・グリッドは、一箇所に存在するか、または広い範囲に分散する場合がある。本明細書では第一に、純粹にローカル・エリア技法で構成される、単一構内規模のネットワーク状況におけるコンピューティング・グリッドについて説明する。次いで本明細書では、コンピューティング・グリッドがワイド・エリア・ネットワーク(WAN)を介して分散された場合について説明する。

【0031】図2は、ローカル・コンピューティング・グリッド208を含む、拡張可能コンピューティング・システム200の一構成を示す、構成図である。本明細書では、「拡張可能」とは一般に、システムがフレキシブルかつスケラブルであり、特定企業またはユーザに対してオン・デマンドでコンピューティング能力を増減する機能を有するという意味である。ローカル・コンピューティング・グリッド208は、多数のコンピューティング要素CPU1、CPU2、・・・CPU_nで構成される。実施形態の一例では、10,000またはそれ以上のコンピューティング要素が存在可能である。これらのコンピューティン

グ要素は、長寿命の状態情報を要素ごとを含むかまたは記憶することがないため、ローカル・ディスクなどの持続的または不揮発性の記憶装置なしに構成することができる。代わりに、すべての長寿命状態情報は、コンピューティング要素とは別の、1つまたは複数のSANスイッチ202を含む記憶域ネットワーク(SAN)を介してコンピューティング要素に結合された、ディスクDISK1、DISK2、・・・DISK_n上に記憶される。好適なSANスイッチは、Brocade (TM) およびExcel (TM) から市販されている。

【0032】すべてのコンピューティング要素は、仮想LAN(VLAN)に分けることのできる1つまたは複数のVLANスイッチ204を介して互いに相互接続されている。VLANスイッチ204はインターネット106に結合される。一般に、コンピューティング要素には、VLANスイッチに接続された1つまたは2つのネットワーク・インターフェースが含まれる。図2では、簡略化するために、すべてのノードが2つのネットワーク・インターフェースを備えるように示されているが、これよりも少ないかまたは多いネットワーク・インターフェースを備えることも可能である。現在では、多くの製造販売業者が、VLAN機能をサポートするスイッチを提供している。たとえば、好適なVLANスイッチは、Cisco Systems, Inc. (TM) およびXtreme Networks (TM) から市販されている。同様に、ファイバ・チャネル・スイッチ、SCSI対ファイバ・チャネル・ブリッジング・デバイス、およびネットワーク結合記憶装置(NAS/Network Attached Storage) デバイスを含み、SANを構築するための製品が数多く市販されている。

【0033】制御プレーン206は、SAN制御バス、CPU制御バス、およびVLAN制御バスによって、それぞれ、SANスイッチ202、CPU1、CPU2、・・・CPU_nといったCPU、およびVLANスイッチ204に結合される。

【0034】それぞれのVSFは、VLANセット、VLANに取り付けられたコンピューティング要素セット、およびコンピューティング要素セットに結合されたSAN上で使用可能な記憶装置のサブセットで構成される。SAN上

で使用可能な記憶装置のサブセットはS A Nゾーンと呼ばれ、S A Nハードウェアによって他のS A Nゾーンの一部分であるコンピューティング要素からのアクセスから保護される。偽造不可能なポート識別子を提供するV L A Nを使用して、一顧客またはエンド・ユーザが、他の顧客またはエンド・ユーザのV S F資源へのアクセスを取得しないようにすることが好ましい。

【0035】図3は、S A Nゾーンを特徴とする仮想サーバ・ファームの一例を示す構成図である。複数のWe bサーバW S 1、W S 2、などが、第1のV L A N (V L A N 1) によって、ロード・バランサ (L B) /ファイアウォール302に結合される。第2のV L A N (V L A N 2) は、インターネット106とロード・バランサ (L B) /ファイアウォール302とを結合する。各We bサーバは、本明細書でさらに説明するメカニズムを使用して、C P U 1、C P U 2、などの中から選択することができる。We bサーバはS A Nゾーン304に結合され、これは1つまたは複数の記憶デバイス306 a、306 bに結合される。

【0036】任意の所与の時点で、図2のC P U 1などのコンピューティング・グリッド内のコンピューティング要素は、単一のV S Fに関連付けられたV L A NのセットおよびS A Nゾーンに接続されているだけである。V S Fは典型的には、異なる組織間で共用されない。S A N上にあり、単一のS A Nゾーンに属した記憶装置のサブセット、ならびにこれとV L A N上にあるコンピューティング要素とに関連付けられたV L A Nセットが、V S Fを定義する。

【0037】V L A NのメンバシップおよびS A Nゾーンのメンバシップを制御することによって、制御プレーンは、コンピューティング・グリッドを複数のV S Fに論理的に区分させる。1つのV S Fのメンバは、他のV S Fのコンピューティング資源または記憶資源にアクセスすることができない。こうしたアクセス制約は、V L A Nスイッチによってハードウェア・レベルで、ならびにファイバ・チャネル・スイッチなどのS A NハードウェアおよびS C S I対ファイバ・チャネル・ブリッジング・ハードウェアなどのエッジ・デバイスのポート・レベルのアクセス制御メカニズム（たとえばゾーニング）によって、強制される。コンピューティング・グリッドの一部を形成するコンピューティング要素は、

物理的にはVLANスイッチおよびSANスイッチの制御ポートまたはインターフェースに接続されていないため、VLANまたはSANゾーンのメンバシップを制御することができない。したがって、コンピューティング・グリッドのコンピューティング要素は、それらが格納されているVSF内にはないコンピューティング要素にはアクセスすることができない。

【0038】制御プレーンを実行するコンピューティング要素だけが、グリッド内にあるデバイスの制御ポートまたはインターフェースに物理的に接続されている。コンピューティング・グリッド内にあるデバイス（コンピュータ、SANスイッチ、およびVLANスイッチ）は、こうした制御ポートまたはインターフェースを介してのみ構成することができる。これによって、コンピュータ・グリッドを複数のVSFに動的に区分させる、単純ではあるが非常に確実な手段が提供される。

【0039】VSF内にあるそれぞれのコンピューティング要素は、任意の他のコンピューティング要素に置き換えられる。所与のVSFに関連付けられたコンピューティング要素、VLAN、およびSANゾーンの数、時間の経過と共に、制御プレーンの制御の下で変化することがある。

【0040】一実施形態では、コンピューティング・グリッドには、予備にとってある多数のコンピューティング要素を含むアイドル・プールが含まれる。アイドル・プールからのコンピューティング要素は、VSFが利用可能なCPUまたはメモリ容量を増やすため、またはVSF内にある特定のコンピューティング要素の障害を処理するために、特定のVSFに割り当てることができる。コンピューティング要素がWebサーバとして構成される場合、アイドル・プールは、変化するかまたは「満杯の」Webトラフィック負荷および関連するピーク時処理負荷のための、大規模な「緩衝器」としての働きをする。

【0041】アイドル・プールは、多くの異なる組織間で共用されるため、単一の組織がアイドル・プールの全経費を支払う必要がなくなり、スケール・メリットが与えられる。異なる組織が1日のうちで異なる時間に必要に応じてアイドル・プールからコンピューティング要素を取得することが可能であり、それによって、各VSFを必要なときに展開させ、トラフィックが通常のレベルに落

ち着いたときには縮小させることができる。多くの異なる組織が同時にピークの状態を持続し、それによって潜在的にアイドル・プールの容量を使い尽くしてしまう場合、CPUおよび記憶要素を追加することでアイドル・プールを増やすことができる（スケーラビリティ）。アイドル・プールの容量は、定常状態で、特定のVSFが必要なときにアイドル・プールから追加のコンピューティング要素を取得できない確率を大幅に減らすように設計されている。

【0042】図4A、図4B、図4C、および図4Dは、コンピューティング要素をアイドル・プールの内外へ移動させることに関連した、連続するステップを示す構成図である。はじめに図4Aを参照すると、制御プレーンが、VSF1、VSF2と表示された第1および第2のVSFに論理的に接続されたコンピューティング・グリッドの要素を有すると想定される。アイドル・プール400には複数のCPU402が含まれ、そのうちの1つがCPUXと表示されている。図4Bでは、VSF1に追加のコンピューティング要素が必要であることが明らかになった。したがって、経路404で示されるように、制御プレーンはCPUXをアイドル・プール400からVSF1に移動させる。

【0043】図4Cでは、VSF1にはCPUXが不要になったため、制御プレーンがCPUXをVSF1からアイドル・プール400へ戻すように移動させる。図4Dでは、VSF2に追加のコンピューティング要素が必要であることが明らかになった。したがって、制御プレーンはCPUXをアイドル・プール400からVSF2に移動させる。このようにして、時間の経過と共にトラフィックの条件が変化するに従って、単一のコンピューティング要素がアイドル・プールに属し（図4A）、その後特定のVSFに割り当てられ（図4B）、その後アイドル・プールに戻され（図4C）、その後他のVSFに属する（図4D）ことが可能である。

【0044】これらの各段階で、制御プレーンは、コンピューティング要素に関連付けられたLANスイッチおよびSANスイッチが、特定のVSF（またはアイドル・プール）に関連付けられたVLANおよびSANゾーンの一部となるように構成する。一実施形態によれば、それぞれの移行の間に、コンピューティング要素は電源を切るかまたはリブートされる。再度電源を入れると、オペ

レーティング・システム（たとえばLinux（TM）、NT（TM）、Solaris（TM）など）のブート可能イメージを含む、SAN上にある別の部分の記憶ゾーンを表示する。記憶ゾーンには、各組織に特有のデータ部分も含まれる（たとえばWebサーバに関連付けられたファイル、データベース区画など）。これは、他のVSFのVLANセットの一部である他のVLANの一部でもあるため、移行先であるVSFのVLANに関連付けられた、CPU、SAN記憶デバイス、およびNASデバイスにアクセスすることができる。

【0045】好ましい実施形態では、記憶ゾーンには、コンピューティング要素によって想定可能な役割に関連付けられた、複数の事前に定義された論理設計図が含まれる。初期時には、Webサーバ、アプリケーション・サーバ、データベース・サーバなどの、任意の特定の役割またはタスク専用のコンピューティング要素はない。コンピューティング要素の役割は、複数の事前に定義され記憶された設計図の1つから取得されるものであって、設計図はそれぞれが、役割に関連付けられたコンピューティング要素にブート・イメージを定義するものである。設計図は、ブート・イメージ位置を役割に関連付けることができる、ファイル、データベース・テーブル、または任意の他の記憶様式の形態で格納することができる。

【0046】したがって、図4A、図4B、図4C、および図4DでのCPUXの移動は物理的ではなく論理的であって、制御プレーンの制御下でVLANスイッチおよびSANゾーンを再構成することによって達成される。さらに、初期にコンピューティング・グリッド内にある各コンピューティング要素は、本来は代替可能なものであり、仮想サーバ・ファーム内で接続された後にのみ特有の処理に関する役割を想定して、ブート・イメージからソフトウェアをロードする。Webサーバ、アプリケーション・サーバ、データベース・サーバなどのような、任意の特定の役割またはタスク専用のコンピューティング要素はない。コンピューティング要素の役割は、複数の事前に定義され記憶された設計図の1つから取得されるものであって、設計図はそれぞれが、役割に関連付けられたコンピューティング要素にブート・イメージを定義するものである。

【0047】任意の所与のコンピューティング要素（ローカル・ディスク

など)に記憶された長寿命の状態情報がないため、異なるV S F間でノードを容易に移動させることが可能であり、まったく異なるO Sおよびアプリケーション・ソフトウェアを実行することができる。さらにこれによって、予定されているかまたは予定外のダウンタイム発生時に、各コンピューティング要素が非常に交換しやすくなる。

【0048】特定のコンピューティング要素は、様々なV S Fから出し入れされる場合に、異なる役割を果たすことができる。たとえば、コンピューティング要素は、1つのV S FにおいてWebサーバとして動作することが可能であり、異なるV S Fに入れられたときにはデータベース・サーバ、Webロード・バランサ、ファイアウォールなどとなることができる。また、異なるV S F内でLinux (TM)、NT (TM)、またはSolaris (TM)などの異なるオペレーティング・システムを連続的にブートして実行することもできる。したがって、コンピューティング・グリッド内にある各コンピューティング要素は代替可能であり、固定的な役割は与えられていない。そのため、コンピューティング・グリッドの予備的処理能力全体を使用して、任意のV S Fが必要とする任意のサービスを提供することができる。これにより、特定のサービスを実行している各サーバが、潜在的に同じサービスを提供できる何千ものバックアップ・サーバを備えることになるため、単一のV S Fが提供するサービスに対して、高い使用可能度および信頼性が与えられる。

【0049】さらに、コンピューティング・グリッドの大規模な予備的処理能力によって、動的なロード・バランシング特性、ならびに高いプロセッサ使用可能度の両方を提供することができる。この能力は、V L A Nを介して相互接続され、S A Nを介して記憶デバイスの構成可能ゾーンに接続された、すべてが制御プレーンによってリアルタイムで制御される、ディスクのないコンピューティング要素を固有に組み合わせることによって実行することができる。あらゆるコンピューティング要素は、任意のV S F内で任意の必須サーバの役割を果たすことが可能であり、S A N内の任意のディスクの任意の論理区画に接続可能である。グリッドがさらに多くのコンピューティング能力またはディスク容量を必要とする場合、コンピューティング要素またはディスク記憶域がアイドル・プール

に手動で追加され、これはより多くの組織にV S Fサービスが提供されるにつれて、経時的に減らすことができる。C P Uの数、ネットワークおよびディスクの帯域幅、ならびにV S Fが使用可能な記憶域を増やすために、手動介入は不要である。こうしたすべての資源は、アイドル・プール内で制御プレーンが使用可能なC P U、ネットワークおよびディスクの資源から、オン・デマンドで割り振られる。

【0050】特定のV S Fは、手動再構成の対象ではない。アイドル・プール内のマシンのみが、手動でコンピューティング・グリッド内に組み入れられる。その結果、現在手動で構築されたサーバ・ファーム内に存在する潜在的に高い危険性が除去される。新しいサーバを活動中のサーバ・ファーム内に組み入れる際に、人的な誤りによってサーバ・ファームが誤動作を起こす場合があり、その結果、そのW e bサイトのユーザに対するサービスが失われることがある、という可能性は仮想的には消去される。

【0051】制御プレーンは、記憶デバイスに接続されたS A Nに格納されたデータの複製も行うため、任意の特定記憶要素に障害が発生しても、システムのどんな部分にもサービスの損失は発生させない。S A Nを使用してコンピューティング・デバイスから長寿命記憶装置を分離し、冗長な記憶装置およびコンピューティング要素を提供することにより、任意のコンピューティング要素を任意の記憶区画に接続することが可能であり、高い使用可能度が達成される。

【0052】

(仮想サーバ・ファームの確立、そのファームへの処理装置の追加、およびそのファームからの処理装置の除去に関する詳細な例)

図5 Aは、実施形態によるV S Fシステムを示す構成図である。図5 Aを参照しながら、V S Fの作成、そのV S Fへのノードの追加、およびそのV S Fからのノードの削除に使用することができる、詳細なステップについて下記で説明する。

【0053】図5 Aは、V L A N対応スイッチ504に結合されたコンピュータA～Gを含む、コンピューティング要素502を示す図である。V L A Nスイッチ504はインターネット106に結合され、V L A Nスイッチにはポー

トV1、V2などがある。さらにコンピュータA～Gは、SANスイッチ506に結合され、これが複数の記憶デバイスまたはディスクD1～D5に結合される。SANスイッチ506にはポートS1、S2などがある。制御プレーン・マシン508は、制御パスおよびデータ・パスによって、SANスイッチ506およびVLANスイッチ504に結合される。制御プレーンは、制御ポートを介してこれらのデバイスに制御コマンドを送信することができる。

【0054】説明を簡単にするために、図5Aではコンピューティング要素の数を少なくしてある。実際には、たとえば何千またはそれ以上に上る多数のコンピュータと、同様に多数の記憶デバイスによって、コンピューティング・グリッドが形成されている。こうした大規模な構造では、複数のSANスイッチが相互接続されてメッシュを形成し、複数のVLANスイッチが相互接続されてVLANメッシュを形成する。ただし、単純でわかりやすくするために、図5Aには単一のSANスイッチおよび単一のVLANスイッチが示されている。

【0055】初期時に、制御プレーンがVSFの作成要求を受け取るまでは、すべてのコンピュータA～Gがアイドル・プールに割り当てられる。VLANスイッチのすべてのポートは特定のVLANに割り当てられ、これをVLAN 1（アイドル・ゾーン用）と表示するものとする。制御プレーンが、1つのロード・バランサ／ファイアウォールと、SAN上の記憶装置に接続された2つのWebサーバを含む、VSFの構築を要求されたものと想定する。制御プレーンへの要求は、管理インターフェースまたは他のコンピューティング要素を介して着信することができる。

【0056】これに応答して、制御プレーンはCPU Aをロード・バランサ／ファイアウォールとして割り当てるかまたは割り振り、CPU BおよびCをWebサーバとして割り振る。CPU Aは、論理上、SANゾーン1に配置され、ロード・バランシング／ファイアウォール用の専用ソフトウェアを含むディスク上のブート可能区画が指し示される。「指し示される」という用語は便宜上使用されており、CPU Aが動作させる必要のある適切なソフトウェアを取得するかまたは見つけることができるのに十分な情報が、何らかの手段によってCPU Aに与えられることを示す意図の用語である。CPU AをSANゾ

ーン1に配置することで、そのSANゾーンのSANによって制御されるディスクからCPU Aが資源を取得することができる。

【0057】ロード・バランサは、CPU BおよびCを2つのWebサーバとして理解するように制御プレーンによって構成され、ロード・バランシングを行うことが想定される。ファイアウォール構成により、CPU BおよびCがインターネット106から許可なくアクセスされるのを防ぐことができる。CPU BおよびCには、特定のオペレーティング・システム（たとえば、Solaris (TM)、Linux (TM)、NT (TM) など）およびWebサーバ・アプリケーション・ソフトウェア（たとえばApache）用のブート可能OSイメージを含む、SAN上のディスク区画が指し示される。VLANスイッチは、VLAN 1上にポートv1およびv2を配置し、VLAN 2上にポートv3、v4、v5、v6、およびv7を配置するよう構成される。制御プレーンは、SANゾーン1にファイバ・チャネル・スイッチ・ポートs2、s3、およびs8を配置するように、SANスイッチ506を構成する。

【0058】CPUにどのように特定のディスク・ドライブが指し示されるか、およびこれがブートアップおよびディスク・データへの共用アクセスに関して何を意味するかについて、さらに説明する。

【0059】図6は、結果として生じる、集合的にVSF1と呼ばれるコンピューティング要素の論理接続を示す構成図である。ディスク・ドライブDD1は、記憶デバイスD1、D2などの中から選択される。図6に示されるように、いったん論理構造が達成されると、CPU A、B、Cに起動コマンドが与えられる。これに応答して、CPU Aは専用のロード・バランサ/ファイアウォール・マシンとなり、CPU B、CはWebサーバとなる。

【0060】次に、ポリシー・ベースの規則により、制御プレーンが、VSF1内に他のWebサーバが必要であると判断するものと想定する。これは、Webサイトに入ってくる要求の数が増えるためであり、顧客プランでは少なくとも3つのWebサーバをVSF1に追加することが可能である。あるいは、VSFを所有または操作する組織が他のサーバを要求し、それ自体のVSFにサーバを追加できるようにする特権が付与されたWebページなどの管理メカニ

ズムを介して追加したものである。

【0061】これに応答して、制御プレーンはCPU DをVSF 1に追加することを決定する。そのために、制御プレーンは、VLAN 2にポートv 8およびv 9を追加して、CPU DをVLAN 2に追加することになる。また、CPU DのSANポートs 4もSANゾーン1に追加される。CPU Dには、Webサーバとしてブートアップし実行するSAN記憶装置のブート可能部分が指し示される。CPU Dは、Webページ・コンテンツ、実行可能サーバ・スクリプトなどからなる、SAN上にある共用データへの読み取り専用アクセスも取得する。この方法で、サーバ・ファームに関するWeb要求が処理可能となり、CPU BおよびCが要求を処理する。さらに制御プレーンは、ロード・バランシングが行われるサーバ・セットの一部としてCPU Dを含むように、ロード・バランサ（CPU A）を構成することにもなる。

【0062】次にCPU Dがブートアップされ、この時点でVSFのサイズが3つのWebサーバと1つのロード・バランサに増やされている。図7は、結果的に生じる論理接続を示す構成図である。

【0063】ここで制御プレーンが、VSF 2と命名されることになり、2つのWebサーバおよび1つのロード・バランサ／ファイアウォールを必要とする、他のVSFを作成する要求を受け取ると想定する。制御プレーンは、CPU Eをロード・バランサ／ファイアウォールとし、CPU F、GをWebサーバとするように割り振る。これでCPU EはCPU F、Gが互いにロード・バランシングされる2つのマシンであることを理解するように構成される。

【0064】この構成を実施するために、制御プレーンは、VLAN 1がポートv 10、v 11を含み（すなわち、インターネット106に接続され）、VLAN 3がポートv 12、v 13、およびv 14、v 15を含むように、VLANスイッチ504を構成する。同様に、SANゾーン2がSANポートs 6およびs 7およびs 9を含むように、SANスイッチ506を構成する。このSANゾーンには、CPU Eをロード・バランサとして実行し、CPU FおよびGを、SANゾーン2のディスクD 2に含まれる共用読み取り専用ディスク区画を使用するWebサーバとして実行するのに必要なソフトウェアを含む、記憶

装置が含まれる。

【0065】図8は、結果的に生じる論理接続を示す構成図である。2つのV S F (V S F 1、V S F 2)は同じ物理V L A NスイッチおよびS A Nスイッチを共用するが、2つのV S Fは論理的に区分されている。C P U B、C、Dにアクセスするユーザ、またはV S F 1を所有するかまたは動作させる企業は、V S F 1のC P Uおよび記憶装置にのみアクセスできる。こうしたユーザは、V S F 2のC P Uまたは記憶装置にはアクセスできない。これは、唯一の共用セグメント(V L A N 1)上にある別々のV L A Nおよび2つのファイアウォールと、2つのV S Fが構成される異なるS A Nゾーンとの組み合わせによって生じるものである。

【0066】さらに、その後、制御プレーンはV S F 1を2つのWe bサーバに戻すことを決定すると想定する。これは、V S F 1上で一時的に増加した負荷が減少したか、または何らかの他の管理的処置が講じられたためである。これに応答して、制御プレーンは、C P Uの電源切断を含む可能性のある特別なコマンドによって、C P U Dをシャットダウンする。C P Uがいったんシャットダウンされると、制御プレーンはポートv 8およびv 9をV L A N 2から除去し、S A Nポートs 4もS A Nゾーン1から除去する。ポートs 4はS A Nのアイドル・ゾーンに配置される。S A Nのアイドル・ゾーンには、たとえばS A Nゾーン1(アイドル用)またはゾーン0が指定される。

【0067】その後何らかの時点で、制御プレーンは別のノードをV S F 2に追加するように決定することができる。これは、V S F 2内のWe bサーバにかかる負荷が一時的に増加したか、または他の理由によって生じるものである。したがって、制御プレーンは、破線バス802で示されるように、C P U DをV S F 2に配置することを決定する。そのためには、V L A N 3がポートv 8、v 9を含み、S A Nゾーン2がS A Nポートs 4を含むように、V L A Nスイッチを構成する。C P U Dには、V S F 2のサーバに必要なO SおよびWe bサーバ・ソフトウェアのブート可能イメージを含む、ディスク・デバイス2上の記憶部分が指し示される。またC P U Dには、V S F 2内の他のWe bサーバによって共用されているファイル・システムにあるデータへの読取

り専用アクセス権も与えられる。CPU Dは、再度電源が投入され、VSF 2内でロード・バランシングされたWebサーバとして実行され、その後は、VLAN 2に接続されたSANゾーン1またはCPUにあるどんなデータにもアクセスできない。具体的に言えば、たとえCPU Dが以前はVSF 1の一部であったとしても、VSF 1のどんな要素にもアクセスする手段がなくなる。

【0068】さらに、この構成では、CPU Eが強制するセキュリティ外辺部が、CPU Dを含むように動的に拡張されている。したがって、実施形態では、VSFに追加されるかまたはVSFから除去されるコンピューティング要素を適切に保護するように自動的に調整する、動的なファイアウォールを提供する。

【0069】

(SANのディスク・デバイス)

ブートアップのため、または他のノードと共用する必要があるディスク記憶装置にアクセスするために、CPUに対してSAN上にある特定のデバイスを指し示す方法、あるいはそうでなければブートアップ・プログラムおよびデータを見つけることのできる情報を提供することができる方法がいくつかある。

【0070】その1つが、コンピューティング要素に接続されたSCSI対ファイバ・チャネル・ブリッジング・デバイスを使用し、ローカル・ディスクにSCSIインターフェースを提供する方法である。そのSCSIポートをファイバ・チャネルSAN上の正しいドライブに経路指定することによって、コンピュータは、ローカル接続されたSCSIディスクにアクセスするのと同様に、ファイバ・チャネルSAN上にある記憶デバイスにアクセスすることができる。したがって、ブートアップ・ソフトウェアなどのソフトウェアは、ローカル接続されたSCSIディスクをブートオフするのと同様に、SAN上にあるディスク・デバイスをブートオフするだけである。

【0071】もう1つは、ノードおよび関連するデバイス・ドライバ上にファイバ・チャネル・インターフェースを設け、ファイバ・チャネル・インターフェースをブート・デバイスとして使用できるようにするROMおよびOSソフトウェアをブートする方法である。

【0072】もう1つは、SCSIまたはIDEデバイス制御装置のようであるが、ディスクにアクセスするためにSANを介して通信する、インターフェース・カード（たとえばPCIバスまたはSバス）を設ける方法である。Solaris (TM) およびWindows (R) NTなどのオペレーティング・システムは、全体的に、この代替例で使用可能なディスクなしのブート機能を提供する。

【0073】典型的には、所与のノードに関連付けられたSANディスク・デバイスには2種類ある。その第1は、他のコンピューティング要素と論理的には共用されず、通常はブート可能OSイメージ、ローカル構成ファイルなどを含むノードごとのルート区画を構成する種類のディスクである。これはUnix (R) システムのルート・ファイルシステムと等価である。

【0074】第2の種類のディスクは、他のノードと記憶装置を共用する。この種の共用は、CPU上で実行されるOSソフトウェアおよびノードが共用記憶装置にアクセスする必要性によって変化する。OSが、複数のノード間での共用ディスク区画への読取り／書込みアクセスを可能にする、クラスタ・ファイル・システムを提供する場合、こうしたクラスタ・ファイル・システムとして共用ディスクが取り付けられる。同様に、システムは、クラスタ内で実行中の複数のノードが共用ディスクに同時に読取り／書込みアクセスできるようにする、Oracle Parallel Server (TM) などのデータベース・ソフトウェアを使用することができる。このような場合、共用ディスクはすでに、基本となるOSおよびアプリケーション・ソフトウェアに組み込まれている。

【0075】OSおよび関連するアプリケーションが他のノードと共用のディスク・デバイスを管理できないために、こうした共用アクセスができないオペレーティング・システムの場合、共用ディスクを読取り専用デバイスとして取り付けることができる。多くのWebアプリケーションの場合、Web関連ファイルへの読取り専用アクセスがあれば十分である。たとえば、Unix (R) システムでは、特定のファイルシステムを読み取り専用として取り付けることができる。

【0076】

(多重スイッチ・コンピューティング・グリッド)

図5Aに関連して上記で述べた構成は、大規模交換VLAN構造を形成するために複数のVLANスイッチを相互接続させることによって、また大規模交換SANメッシュを形成するために複数のSANスイッチを相互接続させることによって、多数のコンピューティング・ノードおよび記憶ノードに拡張することができる。この場合、コンピューティング・グリッドは、図4に一般的に示されたアーキテクチャを備えるが、CPUおよび記憶デバイス用の非常に多数のポートを含むSAN/VLAN交換メッシュは除く。制御プレーン上で動作するいくつかのマシンは、以下で説明するように、VLAN/SANスイッチの制御ポートに物理的に接続することができる。当分野では、複数のVLANスイッチを相互接続して、複雑な多重キャンパス内データ・ネットワークを作成することが知られている。たとえば、http://www.cisco.com/warp/public/cc/sol/mkt/ent/ndsgn/highd_wp.htmからオンラインで入手可能な、G. Havilandの「Designing High-Performance Campus Intranets with Multilayer Switching」Cisco Systems, Inc. (TM)を参照されたい。

【0077】

(SANアーキテクチャ)

この説明では、SANがファイバ・チャネル・スイッチおよびディスク・デバイスを備え、潜在的にはSCSI対ファイバ・チャネル・ブリッジなどのファイバ・チャネル・エッジ・デバイスを備えることを想定している。ただし、SANは、ギガビット・イーサネット(R)・スイッチ、または他の物理層プロトコルを使用するスイッチなどの代替技法を使用して構築することも可能である。具体的に言えば、IPを介してSCSIプロトコルを実行することにより、IPネットワークを介してSANを構築するための努力が、現在進行中である。前述の方法およびアーキテクチャは、SANを構築するためのこうした代替方法に適合可能である。VLAN実行可能層2環境を介して、IPを介したSCSIのようなプロトコルを実行することによって、SANが構築される場合、SANゾーンは

、これらを異なるVLANにマッピングすることによって作成される。

【0078】また、高速イーサネット(R)またはギガビット・イーサネット(R)などのLAN技法を介して動作する、ネットワーク結合記憶装置(NAS)も使用可能である。このオプションでは、コンピューティング・グリッドのセキュリティおよび論理区分を実施するために、SANゾーンの代わりに異なるVLANが使用される。こうしたNASデバイスは、典型的には、複数のノードが同じ記憶装置を共用できるようにするために、Sun(TM)のNSFプロトコルまたはMicrosoft(TM)のSMBなどのネットワーク・ファイルシステムをサポートする。

【0079】

(制御プレーンの実施)

前述の説明では、制御プレーンはSAN/VLANスイッチの制御ポートおよびデータ・ポートに結合されたボックスとして表されている。ただし、他の制御プレーンの実施も検討されている。

【0080】典型的には、SAN/VLAN制御ポートはイーサネット(R)・インターフェースである。図9は、こうした場合に使用可能なアーキテクチャを示す構成図である。各VLANスイッチ(VLAN SW1、VLAN SWn)のすべての制御(「CTL」)ポートおよび各SANスイッチ(SAN SW1、SAN SWn)のすべての制御ポートが、単一のイーサネット(R)・サブネット902上に配置される。サブネット902は複数の制御プレーン・マシンCP CPU1、CP CPU2などにのみ接続される。これにより、複数の制御プレーン・マシンを、すべてのSANスイッチおよびVLANスイッチの制御ポートに接続することができる。

【0081】この構成では、複数の制御プレーン・マシンは集合的に、制御プレーンまたはCP 904と呼ばれる。CP 904内のマシンのみが、VLANスイッチおよびSANスイッチの制御ポートへの物理的な接続を有する。したがって、所与のVSFにあるCPUが、独自のVSFまたは任意の他のVSFに関連付けられたVLANおよびSANゾーンのメンバシップを変更することはできない。

【0082】あるいは、イーサネット（R）・インターフェースの代わりに、制御ポートがシリアル・ポートまたはパラレル・ポートであってもよい。この場合、ポートは制御プレーン・マシンに結合される。

【0083】

（制御プレーン・データとVLANとの接続）

制御プレーンを実行するマシンは、VLANスイッチならびにSANスイッチの両方にあるデータ・ポートへのアクセスを有する必要がある。これは、制御プレーンが特定のノードに関するファイルを構成し、現在のCPU負荷、ネットワーク負荷、およびディスク負荷に関するノードからリアルタイム情報を収集することができるようにするために必要である。

【0084】図5Bは、制御プレーン516とデータ・ポートとを接続するための構成を示す、一実施形態の構成図である。一実施形態では、各VSF内のマシンが、制御プレーンのエージェントとして動作しているマシン510に定期的にパケットを送信する。あるいは、制御プレーン・エージェント・マシン510が、VSF内のノードに対して自らのリアルタイム・データについて定期的にポーリングすることができる。その後、制御プレーン・エージェント・マシン510は、VSF内のすべてのノードから集めたデータを、CP 516に送信する。CP 516内の各マシンは、CP LAN 514に結合される。CP LAN 514は、CPファイアウォール512を介して、VLANスイッチ504の特別ポートV17に結合される。これにより、CPには、すべてのVSFにあるノードからすべてのリアルタイム情報を集めるための、スケーラブルで確実な手段が提供される。

【0085】

（制御プレーンとSANデータとの接続）

図10は、制御プレーン・マシンと、複数のSANスイッチを使用する実施形態（「SANメッシュ」）との接続を示す、構成図である。複数の制御プレーン・マシンCP CPU1、CP CPU2などが、制御プレーン・サーバ・フレーム（CP）904を形成する。各制御プレーン・マシンがSANメッシュのポートに結合される。

【0086】制御プレーン・マシンに関連付けられているのが、制御プレーン専用のデータを含むディスク1004に接続された、SANポートS_o、S_pのセットである。ディスク1004は、制御プレーンがログ・ファイル、統計データ、現在の制御プレーン構成情報、および制御プレーンを実施するソフトウェアを維持する領域である、制御プレーン専用記憶ゾーン1002内に論理的に配置される。SANポートS_o、S_pは、制御プレーンSANゾーンの一部分にしかすぎない。ポートS_o、S_pは、いかなる他のSANゾーン上にも決して配置されず、CP 904の一部であるマシンだけが、これらのポートに接続されたディスク1004にアクセスすることができる。

【0087】ポートS₁、S₂、およびS_n、ならびにポートS_oおよびS_pは、制御プレーンSANゾーン内にある。アイドル・ポートまたは任意のVSFのいずれからのコンピューティング要素も、制御プレーンSANゾーンの一部ではない。これによって、制御プレーン専用データが、任意のVSFからアクセスされるのを確実に防ぐことができる。

【0088】特定の制御プレーン・マシンが、図10のVSF 1などの特定のVSFの一部であるディスク区画にアクセスする必要がある場合は、そのVSFに関連付けられたSANゾーン内に配置される。この例では、CP CPU2がVSF 1のディスクにアクセスする必要があるため、CP CPU2に関連付けられたポートs₂がVSF 1のSANゾーンに配置され、これがポートs₁を含む。CP CPUがポートs₁上のディスクにいったんアクセスすると、VSF 1のSANゾーンから除去される。

【0089】同様に、CP CPU1などのマシンがVSF jのディスクにアクセスする必要がある場合は、VSF jに関連付けられたSANゾーンに配置される。その結果、ポートs₂は、ポートs_jを含むゾーンを含む、VSF jに関連付けられたSANゾーンに配置される。CP 1がポートs_jに接続されたディスクにいったんアクセスすると、VSF jに関連付けられたSANゾーンから除去される。

【0090】

(制御プレーンとVLANデータとの接続)

制御プレーン・マシンは、リアルタイム負荷関連情報などの情報をコンピューティング・ノードから集める必要がある。これを実行するために、制御プレーンは、グリッドそれ自体にあるノードとのネットワーク接続を備える必要がある。

【0091】

(ワイド・エリア・コンピューティング・グリッド)

前述のV S Fは、いくつかの方法でWANを介して分散することができる。

【0092】一代替例では、非同期転送モード(ATM)スイッチングに基づくワイド・エリア・バックボーンが可能である。この場合、ATM LANエミュレーション(LANE)標準の一部であるエミュレーテッドLAN(ELAN)を使用して、各ローカル・エリアVLANがワイド・エリア内に拡張される。この方法では、単一のV S Fが、ATM/SONET/OC-12リンクなどのいくつかのワイド・エリア・リンクにまたがることができる。ELANは、ATM WANをまたがって拡張されるVLANの一部となる。

【0093】あるいは、V S FはVPNシステムを使用し、WANをまたがって拡張される。この実施形態では、ネットワークの基礎となる特徴が無関係となり、WANをまたがって2つまたはそれ以上のV S Fを相互接続して、単一の分散V S Fを作成するために、VPNが使用される。

【0094】分散V S F内でデータをローカル・コピーするために、データ・ミラーリング技術を使用することができる。あるいは、SAN対ATMブリッジングまたはSAN対ギガビット・イーサネット(R)・ブリッジングなどの、いくつかのSAN対WANブリッジング技法のうち1つを使用して、SANがWANを介してブリッジングされる。IPネットワークを介して構築されたSANは、IPがこうしたネットワークを介するとうまく動作するため、WANを介して自然に拡張される。

【0095】図11は、WAN接続を介して拡張される複数のV S Fを示す構成図である。San Joseセンタ、New Yorkセンタ、およびLondonセンタが、WAN接続によって結合される。各WAN接続には、上記の様式で、ATM、ELAN、またはVPN接続が含まれる。各センタには、少なくとも1つのV S Fおよび少なくとも1つのアイドル・プールが含まれる。た

たとえば、San Jose センタにはV S F 1 A およびアイドル・プールAがある。この構成では、任意の他のセンタに配置されたV S F へ割り振るかまたは割り当てるために、センタの各アイドル・プールのコンピューティング資源が使用可能である。こうした割り振りまたは割当てが実行されると、V S F はWAN を介して拡張されることになる。

【0096】

(V S F の使用例)

前述の例で説明したV S F アーキテクチャは、Web サーバ・システムに関連して使用することができる。したがって、前述の例は、特定のV S F においてC P U で構築されるWeb サーバ、アプリケーション・サーバ、およびデータベース・サーバに関して述べてきた。ただし、V S F アーキテクチャは、多くの他のコンピューティング状況で、他の種類のサービスを提供するために使用することが可能であり、これはWeb サーバ・システムに限定されるものではない。

【0097】

(コンテンツ分散ネットワークの一部としての分散V S F)

一実施形態では、V S F が、ワイド・エリアV S F を使用してコンテンツ分散ネットワーク(C D N)を提供する。

【0098】C D N は、データの分散キャッシュを実行するキャッシング・サーバのネットワークである。キャッシング・サーバのネットワークは、たとえば、カリフォルニア州San Mateo のInktomi Corporation (TM) から市販されているトラフィック・サーバ(T S) ソフトウェアを使用して実施することができる。T S はクラスタ・ウェア・システムであって、システムは、キャッシング・トラフィック・サーバ・マシンのセットにさらにC P U が追加されると拡大する。したがって、C P U の追加が拡大のメカニズムであるシステムには好適である。

【0099】この構成では、システムは、T S などのキャッシング・ソフトウェアを実行するV S F の部分にさらにC P U を動的に追加することが可能であり、それによって、Web トラフィックがほぼ満杯状態になるまでキャッシュ容量を増加する。その結果C D N は、適応的な方法で、C P U およびI/O帯域

幅を動的に拡大するように構築することができる。

【0100】

(ホストされるイントラネット・アプリケーション用のVSF)

企業資源計画(ERP)、ORM、およびCRMソフトウェアなどのイントラネット・アプリケーションをホストおよび管理されたサービスとして提供することに対して、関心が高まっている。Citrix WinFrame (TM) およびCitrix MetaFrame (TM) などの技術は、企業が、Microsoft (TM) Windows (R) アプリケーションを、Windows (R) CEデバイスまたはWebブラウザなどのシン・クライアントでのサービスとして提供できるようにするものである。VSFは、こうしたアプリケーションをスケーラブルな方法でホストすることができる。

【0101】たとえば、ドイツのSAP Aktiengesellschaft (TM) から市販されているSAP R/3 ERP (TM) ソフトウェアは、企業が、複数のアプリケーションおよびデータベース・サーバを使用してロード・バランシングを行えるようにするものである。VSFの場合、企業は、リアルタイム要求または他の要素に基づいてVSFを拡大するために、VSFにより多くのアプリケーション・サーバ(たとえばSAPダイアログ・サーバ)を動的に追加する。

【0102】同様に、Citrix Metaframe (TM) は、企業が、より多くのCitrix (TM) サーバを追加することにより、ホストされたWindows (R) アプリケーションを実行しているサーバ・ファーム上のWindows (R) アプリケーション・ユーザを拡大できるようにするものである。この場合、VSFでは、Citrix MetaFrame (TM) VSFはMetaframe (TM) によってホストされたWindows (R) アプリケーションのより多くのユーザに対処するために、動的により多くのCitrix (TM) サーバを追加することになる。

【0103】多くの他のアプリケーションを、前述の例示的な例と同様の方法でホストできることは明らかであろう。

【0104】

(顧客とV S Fとの対話)

V S Fはオン・デマンドで作成されるため、V S Fを「所有する」V S Fの顧客または組織は、V S Fをカスタマイズするために様々な方法でシステムと対話することができる。たとえば、V S Fは制御プレーンを介して即時に作成および修正されるため、V S Fの顧客には、そのV S F自体を作成および修正するための特権的アクセスを認めることができる。特権的アクセスは、Webページおよびセキュリティ・アプリケーションによって与えられるパスワード認証、トークン・カード認証、ケルベロス交換、または他の適切なセキュリティ要素を使用して与えることができる。

【0105】一実施形態例では、Webページのセットが制御プレーン・マシンまたは別々のサーバによって処理される。Webページは、いくつかの層、特定層内のコンピューティング要素の数、各要素に使用されるハードウェアおよびソフトウェア・プラットフォーム、ならびに、これらのコンピューティング要素上でどのような種類のWebサーバ、アプリケーション・サーバ、またはデータベース・サーバのソフトウェアを事前に構成しなければならないかなどを指定することによって、顧客が特別注文のV S Fを作成できるようにするものである。したがって、顧客には仮想提供コンソールが与えられる。

【0106】顧客またはユーザがこうした提供情報を入力すると、制御プレーンは注文を解析および評価して、実行のために待ち行列に入れる。注文は、適切であるかどうかを確認するために、人間の管理者によって再検討することができる。企業が、要求されたサービスに対する支払いに適した信用枠を有するかどうかを確認するために、企業の信用調査を実行することができる。提供注文が承認されると、制御プレーンは注文に合致するV S Fを構成し、V S Fの1つまたは複数のコンピューティング要素へのルート・アクセスを提供するパスワードを、顧客に戻すことができる。その後、顧客は、V S Fで実行するためのアプリケーションのマスタ・コピーをアップロードすることができる。

【0107】コンピューティング・グリッドをホストする企業が、営利目的の企業である場合、Webページは、クレジット・カード、PO番号、電子チェック、または他の支払い方法などの、支払い関連情報も受け取ることができる。

。

【0108】他の実施形態では、Webページは、顧客がリアルタイム負荷に基づいて、要素の最大数から最小数までの間でV S Fを自動的に拡大および縮小させることなどの、いくつかのV S Fサービス・プランのうちの1つを選択できるようにするものである。顧客は、顧客がWebサーバなどの特定層でのコンピューティング要素の最小数、またはV S Fのサーバ処理能力が最低でなければならない時間枠などのパラメータを変更できるようにする、制御値を有することができる。パラメータは、顧客の請求料金を自動的に調整し、課金ログ・ファイル・エントリを生成する、課金ソフトウェアにリンクさせることができる。

【0109】特権的アクセス・メカニズムを介して、顧客は、1秒あたりの使用量、負荷、ヒット、またはトランザクションに関するリアルタイム情報のレポートを取得し、これを監視して、さらにリアルタイム情報に基づいたV S Fの特徴を調整することができる。

【0110】前述の特徴により、サーバ・ファームを構築するための従来の手作業による方法に比べて、はるかに有利であることは明らかであろう。従来の方法では、ユーザは、様々な方法でサーバを追加しサーバ・ファームを構成する際に、面倒な手順書に目を通さなければ、サーバ・ファームの特性に自動的に影響を与えることができない。

【0111】

(V S Fの課金モデル)

V S Fの性質が動的であることから、コンピューティング・グリッドおよびV S Fをホストする企業は、V S Fのコンピューティング要素および記憶要素の実際の使用量に基づいた、V S Fの課金モデルを使用して、V S Fを所有する顧客に対して、サービス料を請求することができる。一律の課金モデルを使用する必要はない。本明細書で開示されたV S Fアーキテクチャおよび方法は、所与のV S Fの資源が静的に割り当てられないため、「従量料金制」の課金モデルを可能にすることができる。したがって、サーバ・ファームにかかる使用負荷がかなり変動的な特定の顧客の場合、一定のピーク時のサーバ処理能力に関連付けられた料金ではなく、実行平均使用量、瞬間使用量などを反映する料金で課金されるた

め、資金を節約することができる。

【0112】たとえば、企業は、サーバ10台など、最低数のコンピューティング要素に対して一定料金を規定し、さらにリアルタイム負荷により10台より多くの要素が必要になった場合には、何台の追加サーバが必要であったか、およびそれが必要であった時間の長さに基づいて、追加サーバに対する増分料金がユーザに課金されることを規定する、課金モデルを使用して、運営することができる。

【0113】こうした請求書の単位は、課金される資源を反映させることができる。たとえば、MIPS時間、CPU時間、何千CPU秒、などの単位で請求書に記載することができる。

【0114】

(顧客の管理下にある制御プレーンAPI)

他の代替例では、資源を変更するための制御プレーンへの呼出しを定義するアプリケーション・プログラミング・インターフェース(API)を顧客に提供することによって、VSFの処理能力を制御することができる。したがって、顧客が作成したアプリケーション・プログラムは、APIを使用してさらに多くのサーバ、多くの記憶装置、多くの帯域幅などを要求するための呼出しまたは要求を発行することができる。この代替例は、顧客がアプリケーション・プログラムにコンピューティング・グリッド環境を認識させ、制御プレーンによって提供される機能を利用させる必要がある場合に、使用することができる。

【0115】上記で開示されたアーキテクチャでは、顧客がコンピューティング・グリッドで使用するために自らのアプリケーションを修正する必要がない。既存のアプリケーションは、手作業で構成されたサーバ・ファームの場合と同様に操作を続行する。ただし、アプリケーションは、制御プレーンが提供するリアルタイム負荷監視機能に基づいて、必要なコンピューティング資源についてよりよく理解すれば、コンピューティング・グリッド内で可能なダイナミズムを利用することが可能である。

【0116】前述の性質を備えたAPIは、アプリケーション・プログラムにサーバ・ファームのコンピューティング能力を変えさせることは可能である

が、既存の手作業による方法を使用してサーバ・ファームを構築することはできない。

【0117】

(自動更新およびバージョン変更)

本明細書で開示した方法およびメカニズムを使用すると、制御プレーンは、VSFのコンピューティング要素で実行されるオペレーティング・システム・ソフトウェアの自動更新およびバージョン変更を実行することができる。したがって、エンド・ユーザまたは顧客は、新しいパッチやバグ修正などを使ってオペレーティング・システムを更新することについて悩む必要がない。制御プレーンは、受け取ったソフトウェア要素のライブラリを維持し、影響を受けるすべてのVSFのコンピューティング要素にこれらを自動的に分散し、インストールすることができる。

【0118】

(実装メカニズム)

コンピューティング要素および監視メカニズムは、いくつかの形式で実装することができる。一実施形態では、各コンピューティング要素は、不揮発性記憶装置1210を除き、図12に示された要素を備えた汎用デジタル・コンピュータであり、監視メカニズムは、本明細書に記載のプロセスを実施するプログラム命令の制御下で動作する、図12に示された種類の汎用デジタル・コンピュータである。

【0119】図12に、本発明の一実施形態を実装し得るコンピュータ・システム1200を描写したブロック図を示す。コンピュータ・システム1200は、情報の通信を行うためのバス1202またはその他の通信メカニズム、およびバス1202に接続された、情報を処理するためのプロセッサ1204を備える。またコンピュータ・システム1200には、バス1202に接続された、情報およびプロセッサ1204によって実行される命令を格納するためのランダム・アクセス・メモリ(RAM)またはその他の動的ストレージ・デバイス等のメイン・メモリ1206が備わる。メイン・メモリ1206は、さらにプロセッサ1204による命令の実行間において、一時変数またはその他の中間情報を格

納するためにも使用される。さらにコンピュータ・システム1200は、プロセッサ1204用の静的な情報ならびに命令を格納するための読み出し専用メモリ（ROM）1208またはその他の静的ストレージ・デバイスを備え、それがバス1202に接続されている。ストレージ・デバイス1210は、磁気ディスクまたは光ディスク等であり、情報および命令を格納するために備えられ、バス1202に接続されている。

【0120】コンピュータ・システム1200には、陰極線管（CRT）等の、コンピュータ・ユーザに情報を表示するためのディスプレイ1212がバス1202を介して接続されることもある。入力デバイス1214は、英数キーおよびその他のキーを備え、バス1202に接続されてプロセッサ1204に情報およびコマンドの選択を伝える。別のタイプのユーザ入力デバイスとして、マウス、トラックボール、またはカーソル移動キー等の、プロセッサ1204に方向情報およびコマンドの選択を伝え、ディスプレイ1212上におけるカーソルの移動をコントロールするためのカーソル・コントロール1216が備わっている。この入力デバイスは、通常、第1の軸（たとえばx軸）および第2の軸（たとえばy軸）からなる2軸に自由度を有しており、それによってこのデバイスは平面内のポジションを指定することができる。

【0121】本発明は、ここに記述した方法、メカニズム及びアーキテクチャを実装するためのコンピュータ・システム1200の使用に関する。本発明の一実施形態によれば、このような方法とメカニズムは、メイン・メモリ1206に収められた1ないしは複数の命令からなる1ないしは複数のシーケンスを実行するプロセッサ1204に応じるシステム1200によって実装される。この種の命令は、ストレージ・デバイス1210等の別のコンピュータ読み取り可能な媒体からメイン・メモリ1206内に読み込んでもよい。メイン・メモリ1206内に収められている命令のシーケンスを実行することにより、プロセッサ1204は、ここに説明したプロセスのステップを実行する。別の実施形態においては、ソフトウェア命令に代えて、あるいはそれと組み合わせてハード・ワイヤード回路を使用し、本発明を実装することもできる。このように本発明の実施態様は、ハードウェア回路およびソフトウェアの特定の組み合わせに限定されるこ

とがない。

【0122】ここで用いている「コンピュータ読み取り可能な媒体」という用語は、プロセッサ1204が実行する命令の提供に与る任意の媒体を指す。その種の媒体は、限定する意図ではないが、不揮発性媒体、揮発性媒体、および伝送媒体を含む各種の形式をとり得る。不揮発性媒体には、たとえば光または磁気ディスクが含まれ、ストレージ・デバイス1210等がこれに該当する。揮発性媒体には、ダイナミック・メモリが含まれ、メイン・メモリ1206等がこれに該当する。伝送媒体には、同軸ケーブル、銅線、および光ファイバが含まれ、バス1202を構成するワイヤーもこれに含まれる。伝送媒体もまた、音波または電磁波、たとえば電波、赤外線、および光データ通信の間に生成される電磁波といった形式をとり得る。

【0123】コンピュータ読み取り可能な媒体の一般的な形態には、たとえば、フロッピー（R）ディスク、フレキシブル・ディスク、ハードディスク、磁気テープ、またはその他の磁気媒体、CD-ROM、その他の光媒体、パンチカード、さん孔テープ、その他孔のパターンを伴う物理的媒体、RAM、PROM、およびEPROM、フラッシュEPROM、その他のメモリ・チップまたはカートリッジ、次に述べる搬送波、またはその他コンピュータによる読み取りが可能な任意の媒体が含まれる。

【0124】各種形式のコンピュータ読み取り可能な媒体が関係して1ないしは複数の命令からなる1ないしは複数のシーケンスがプロセッサ1204に渡され、実行される。たとえば、当初は命令が、リモート・コンピュータの磁気ディスクに収められて運ばれる。リモート・コンピュータは、命令をダイナミック・メモリにロードし、モデムの使用により電話回線を介してその命令を送信することができる。コンピュータ・システム1200に備わるモデムは、電話回線上のデータを受信し、赤外線送信機を使用してそのデータを赤外線信号に変換する。赤外線検出器は、この赤外線信号によって運ばれるデータを受信し、適切な回路がバス1202上にそのデータを乗せる。バス1202は、このデータをメイン・メモリ1206に運び、プロセッサ1204は、そこから命令を取り出して実行する。選択肢の1つとして、プロセッサ1204による実行の前、もしくは

はその後に、メイン・メモリ1206によって受け取られた命令をストレージ・デバイス1210に格納してもよい。

【0125】コンピュータ・システム1200は、通信インターフェース1218も備えており、それがバス1202に接続されている。通信インターフェース1218は、ローカル・ネットワーク1222に接続されるネットワーク・リンク1220に接続されて双方向データ通信を提供する。たとえば、通信インターフェース1218を、対応するタイプの電話回線に接続されてデータ通信を提供する、統合デジタル通信サービス・ネットワーク（ISDN）カードまたはモデムとすることができる。別の例においては、通信インターフェース1218をローカル・エリア・ネットワーク（LAN）カードとし、互換性のあるLANにデータ通信接続を提供することもできる。ワイヤレス・リンクを実装してもよい。この種のいずれの実装においても、通信インターフェース1218は、各種タイプの情報を表すデジタル・データ・ストリームを運ぶ電氣的、電磁氣的、または光学的信号を送受する。

【0126】ネットワーク・リンク1220は、通常、1ないしは複数のネットワークを介して別のデータ・デバイスにデータ通信を提供する。たとえばネットワーク・リンク1220は、ローカル・ネットワーク1222を介してホスト・コンピュータ1224への接続を提供し、あるいはインターネット・サービス・プロバイダ（ISP）1226によって運用されるデータ装置への接続を提供することができる。一方、ISP1226は、現在「インターネット」1228と呼ばれているワールド・ワイド・パケット・データ通信ネットワークを介してデータ通信サービスを提供する。ローカル・ネットワーク1222およびインターネット1228は、いずれもデジタル・データ・ストリームを運ぶ電氣的、電磁氣的、または光学的信号を使用する。これらの各種ネットワークおよびネットワーク・リンク1220上の信号を通り、かつ通信インターフェース1218を通り、コンピュータ・システム1200から、またそこへデジタル・データを運ぶ信号は、情報を伝送する搬送波の一例として挙げた形式である。

【0127】コンピュータ・システム1200は、ネットワーク（1ないしは複数）、ネットワーク・リンク1220および通信インターフェース121

8を介し、プログラム・コードを含めて、メッセージを送信しデータを受信する。インターネットの例においては、サーバ1230がインターネット1228、ISP1226、ローカル・ネットワーク1222および通信インターフェース1218を経由して、要求のあったアプリケーション・プログラム用のコードを送信することが考えられる。本発明に従えば、このようにしてダウンロードしたアプリケーションの1つが、ここに説明された方法及びメカニズムを実装する。

【0128】受信されたコードは、プロセッサ1204によって受信時に実行され、かつ／またはその後に実行するためにストレージ・デバイス1210、あるいはその他の不揮発性ストレージに格納される。このようにしてコンピュータ・システム1200は、搬送波の形式でアプリケーション・コードを獲得することができる。

【0129】

(利点および範囲)

本明細書で開示されたコンピューティング・グリッドは、概念上は、電力グリッドと呼ばれることのある公衆電力網になぞらえることができる。電力グリッドは、単一の大規模電力インフラストラクチャを介して電力サービスを得るために、多くの当事者にスケーラブルな手段を提供する。同様に、本明細書で開示されたコンピューティング・グリッドは、単一の大規模コンピューティング・インフラストラクチャを使用して、多くの組織にコンピューティング・サービスを提供する。電力グリッドを使用する場合、電力消費者は、自分専用の電気機器を独自に管理することはない。たとえば、ユーティリティ消費者にとっては、自分の施設または共用施設で自家発電機を使用し、個人ベースでその容量や増加を管理する理由がない。これに対して、電力グリッドは、膨大な範囲にわたる住民に対して大規模な電力供給が可能であり、これによって大きなスケール・メリットが得られる。同様に、本明細書で開示されたコンピューティング・グリッドは、単一の大規模コンピューティング・インフラストラクチャを使用して、膨大な範囲にわたる住民に対して、コンピューティング・サービスを提供することができる。

【0130】以上、特定の実施形態を参照しながら本発明について述べてきた。ただし、本発明の広範な精神および範囲を逸脱することなく、様々な修正

および変更が実行できることが明らかであろう。したがって、明細書および図面は、限定的なものではなく例示的なものであるとみなされる。

【図面の簡単な説明】

【図1A】単一のマシン・トポロジを有する単純なWebサイトを示す構成図である。

【図1B】1層のWebサーバ・ファームを示す構成図である。

【図1C】3層のWebサーバ・ファームを示す構成図である。

【図2】ローカル・コンピューティング・グリッドを含む拡張可能コンピューティング・システムの1構成を示す構成図である。

【図3】SANゾーンを特徴とする仮想サーバ・ファームの一例を示す構成図である。

【図4A】コンピューティング要素の追加および仮想サーバ・ファームからの要素の除去に関連する連続したステップを示す構成図である。

【図4B】コンピューティング要素の追加および仮想サーバ・ファームからの要素の除去に関連する連続したステップを示す構成図である。

【図4C】コンピューティング要素の追加および仮想サーバ・ファームからの要素の除去に関連する連続したステップを示す構成図である。

【図4D】コンピューティング要素の追加および仮想サーバ・ファームからの要素の除去に関連する連続したステップを示す構成図である。

【図5A】仮想サーバ・ファーム・システム、コンピューティング・グリッド、および監視メカニズムの一実施形態を示す構成図である。

【図5B】監視または制御プレーン・サーバ・ファームがファイアウォールによって保護されているシステムを示す構成図である。

【図6】仮想サーバ・ファームの論理接続を示す構成図である。

【図7】仮想サーバ・ファームの論理接続を示す構成図である。

【図8】仮想サーバ・ファームの論理接続を示す構成図である。

【図9】制御プレーン・サーバ・ファームを示す構成図である。

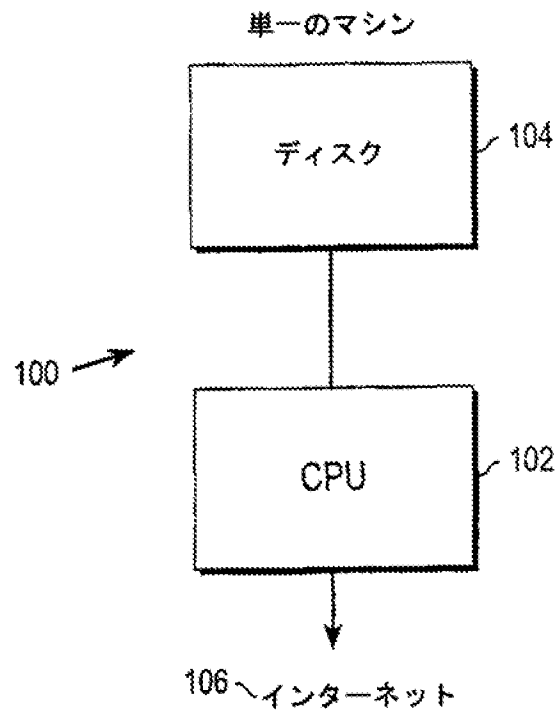
【図10】複数のSANスイッチ（「SANメッシュ」）を使用する実施形態への制御プレーン・マシンの接続を示す構成図である。

【図11】WAN接続を介して拡張される複数のVSFを示す構成図である。

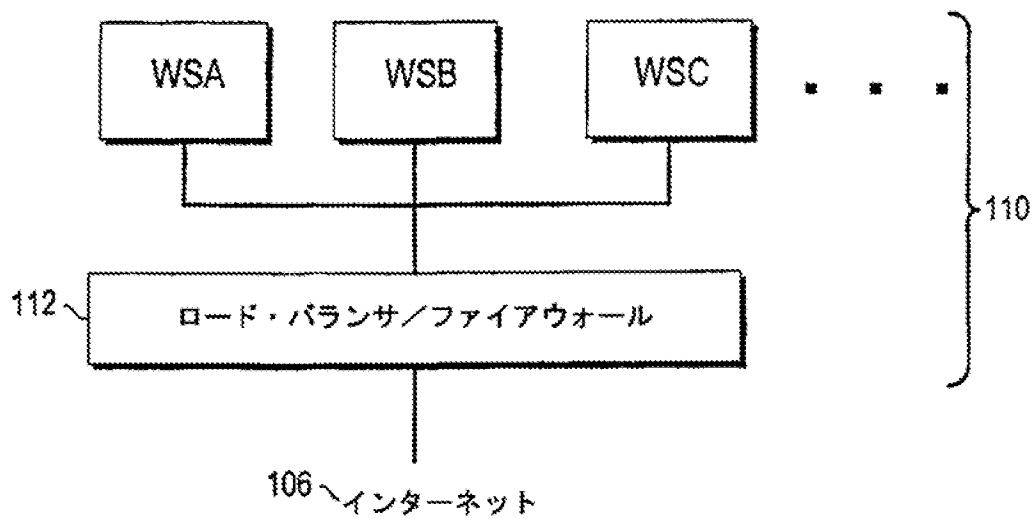
。

【図12】その使用によって実施形態が実施可能なコンピュータ・システムを示す構成図である。

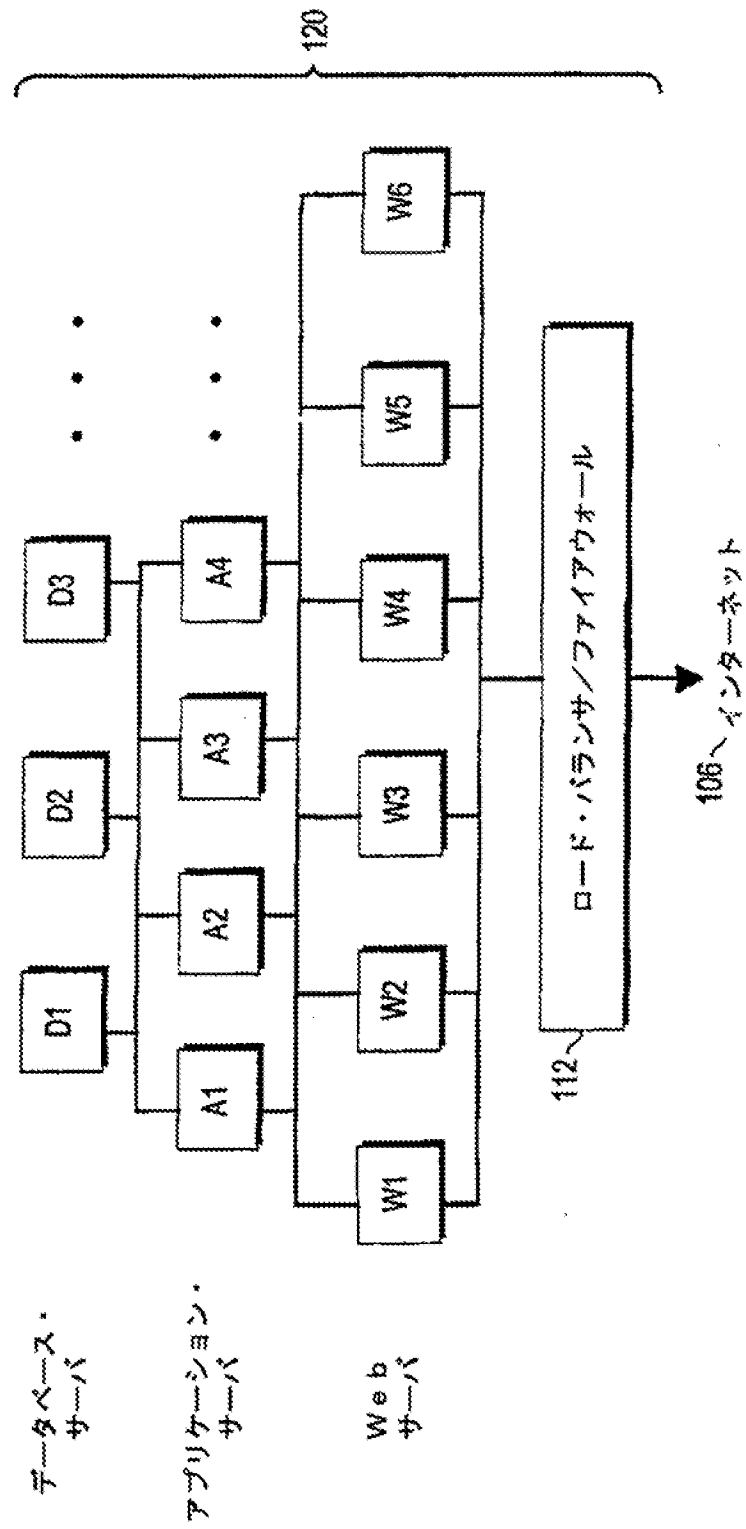
【図1A】



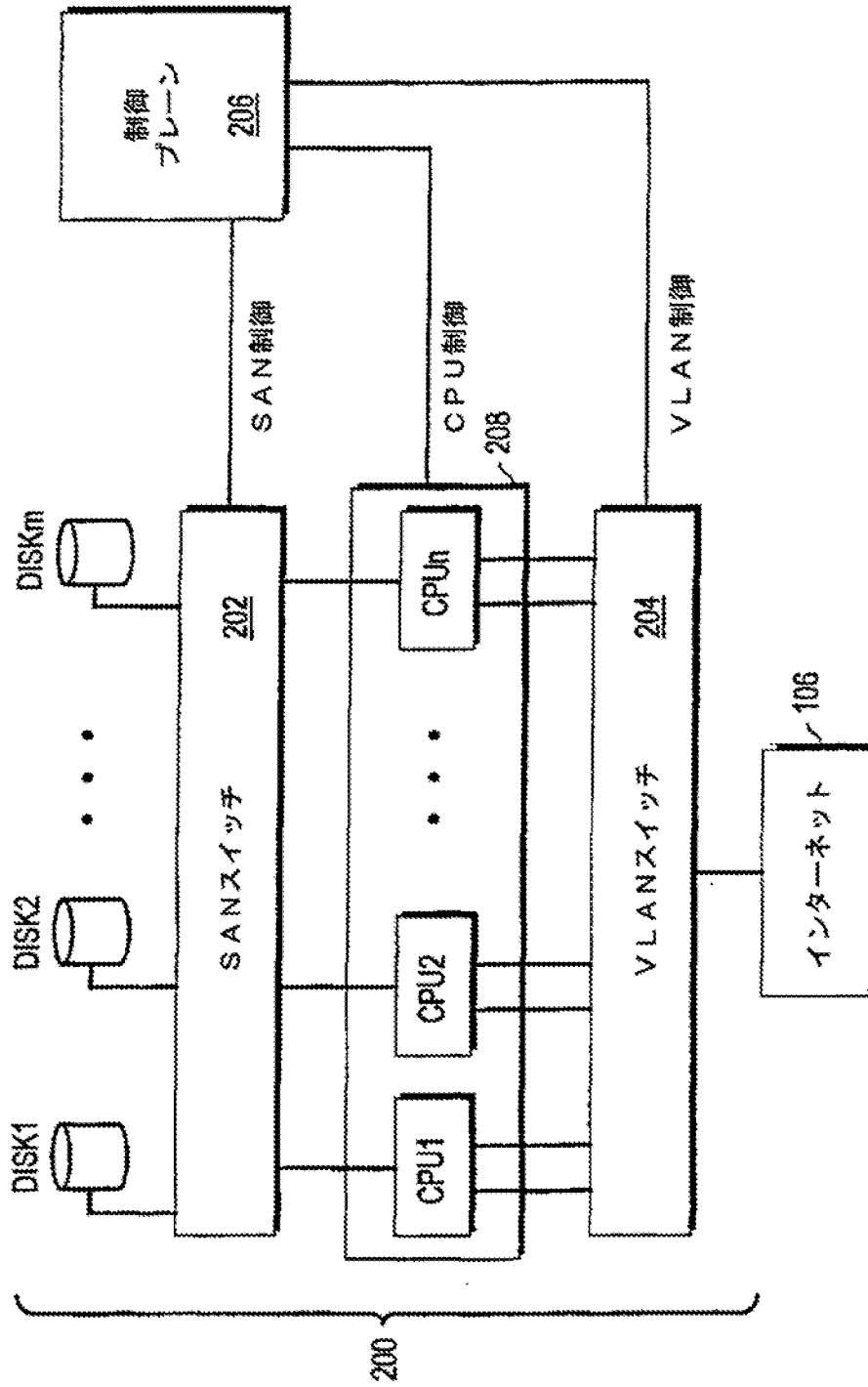
【図1B】



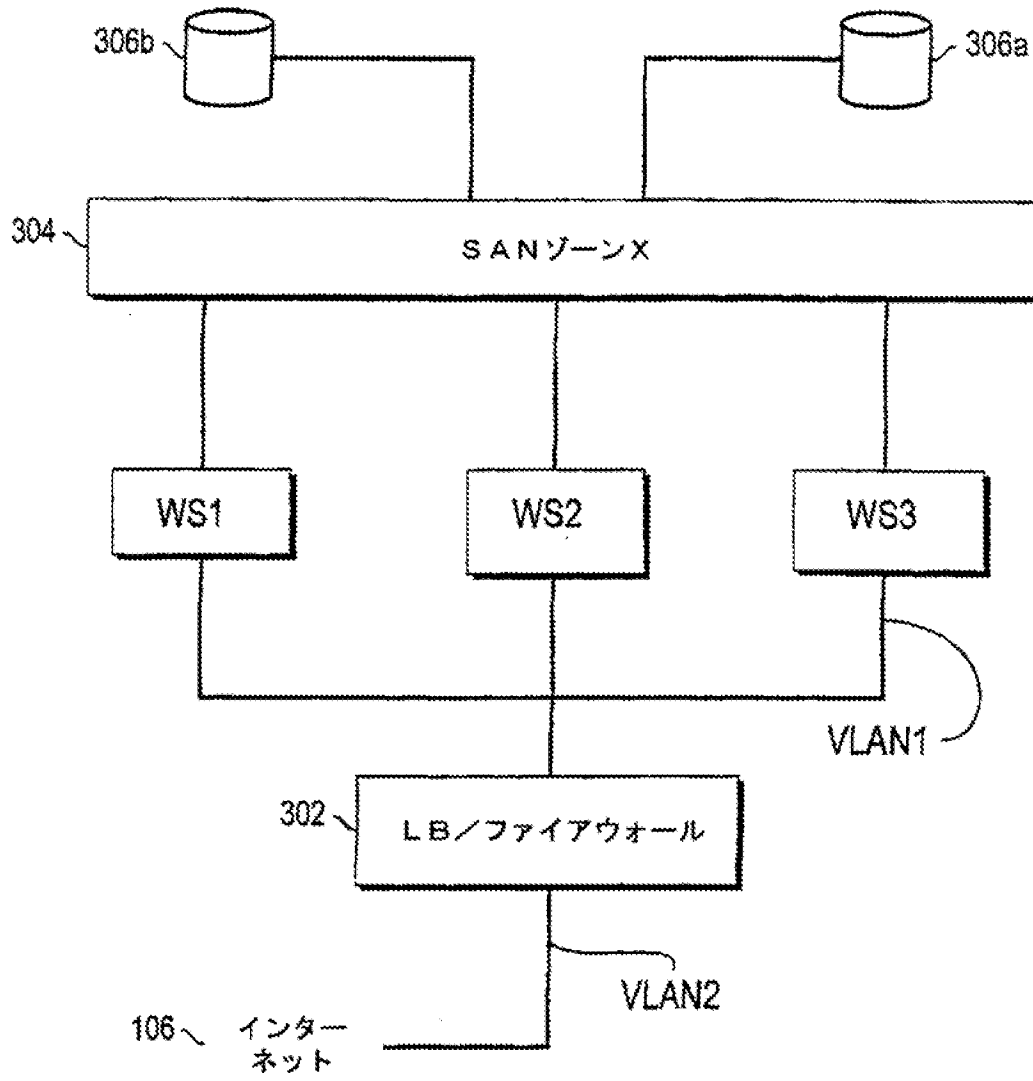
【図1C】



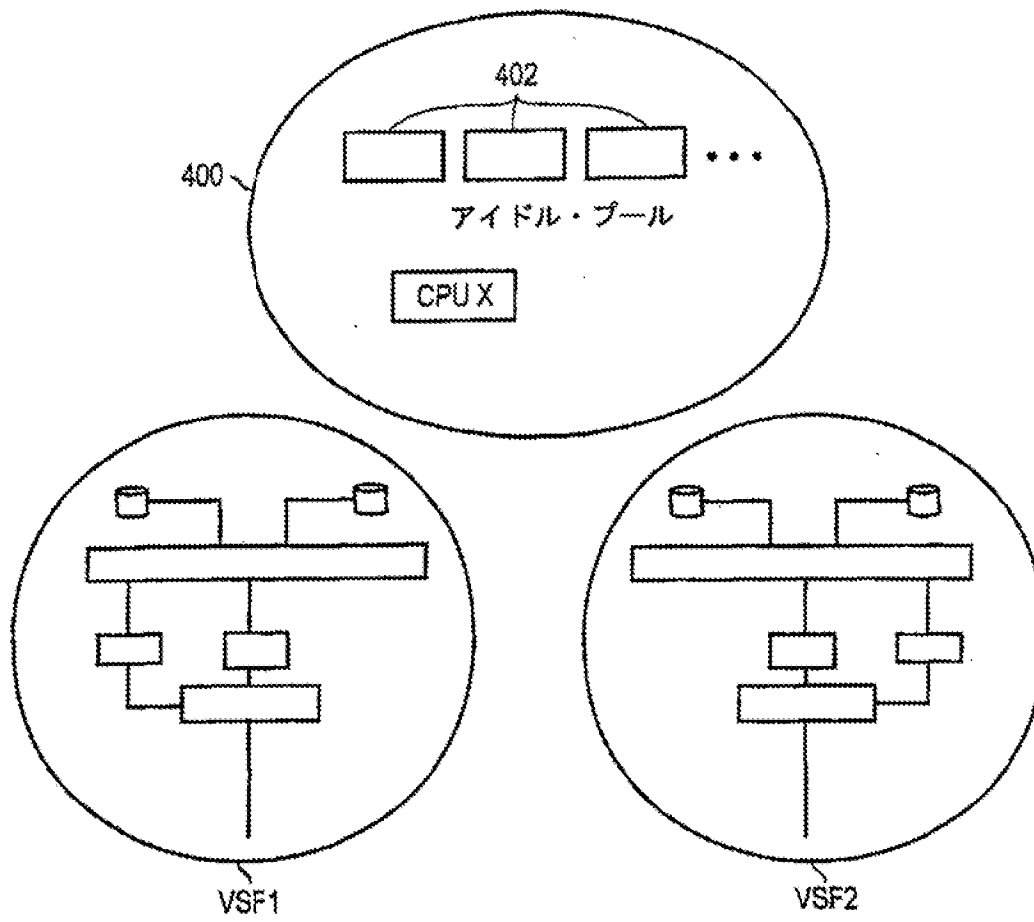
【図2】



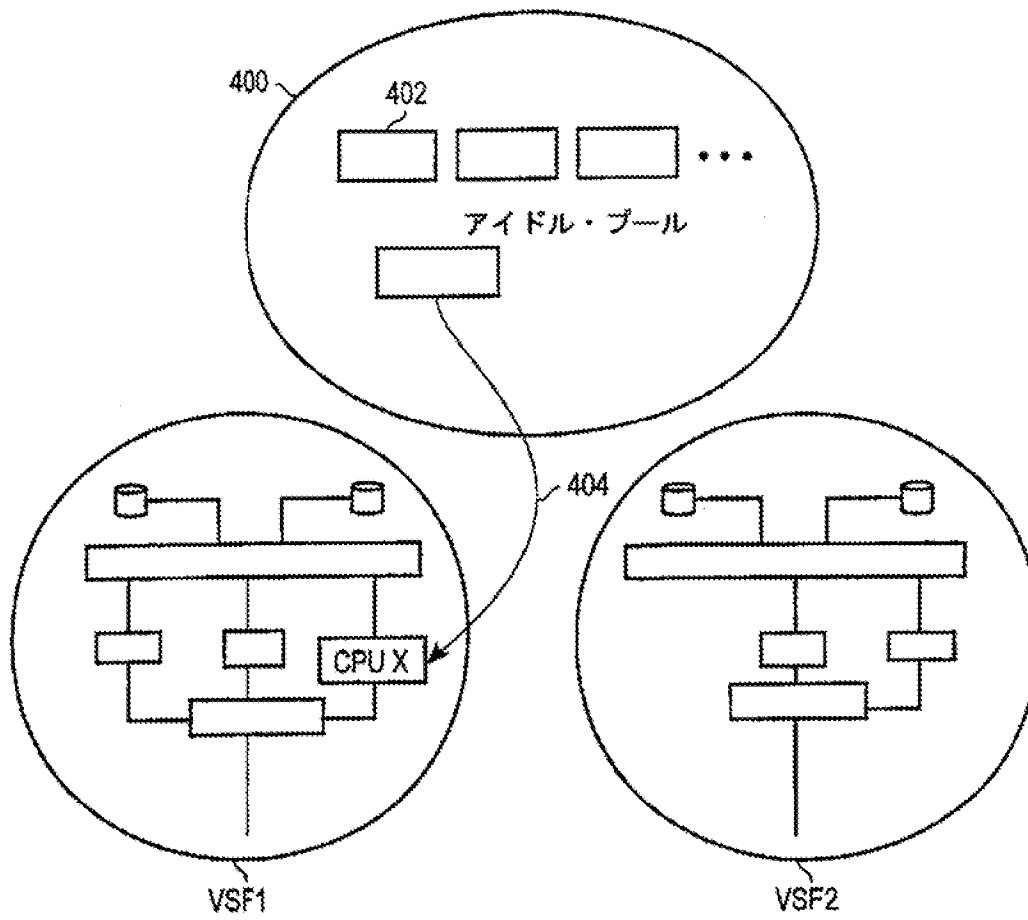
【図3】



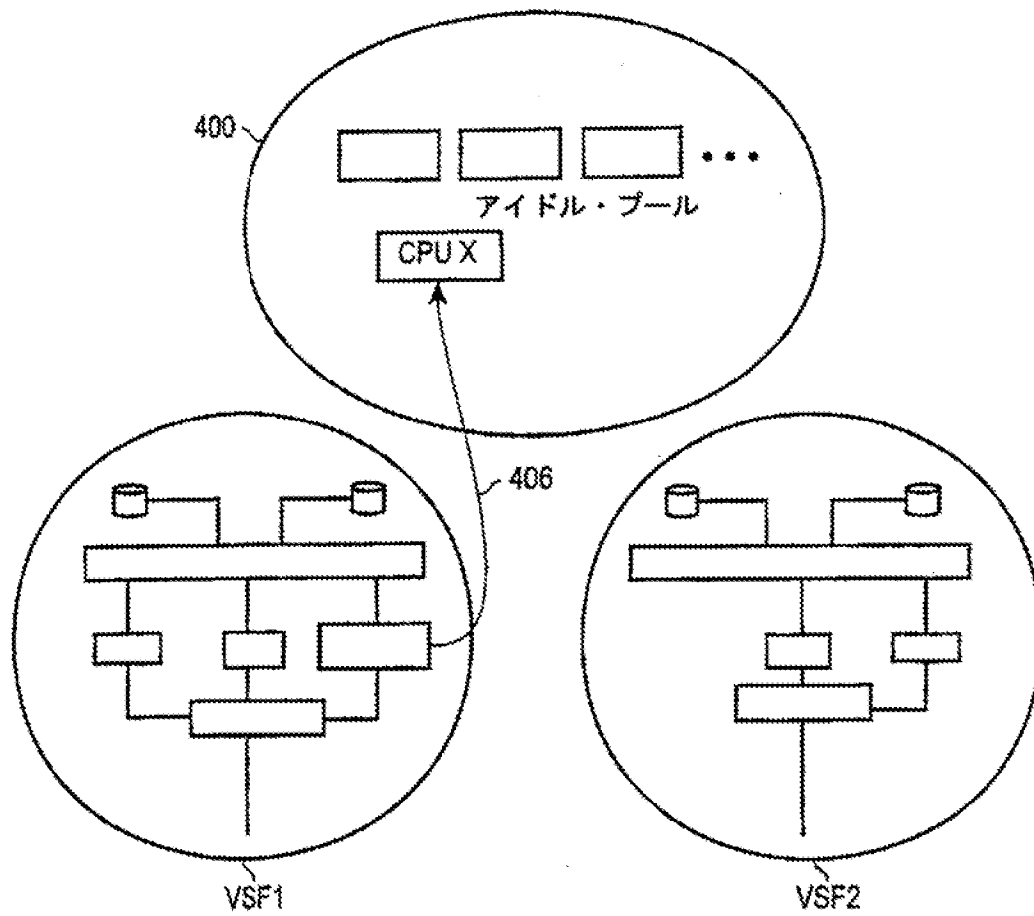
【図4A】



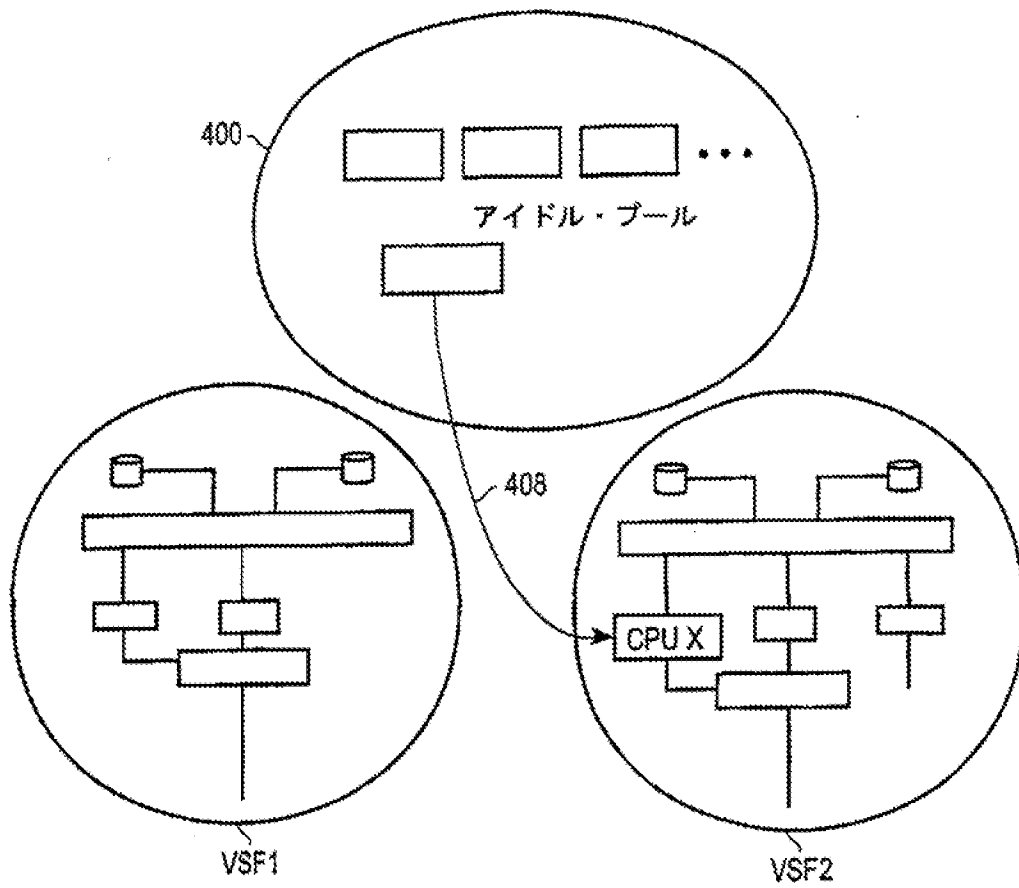
【図4B】



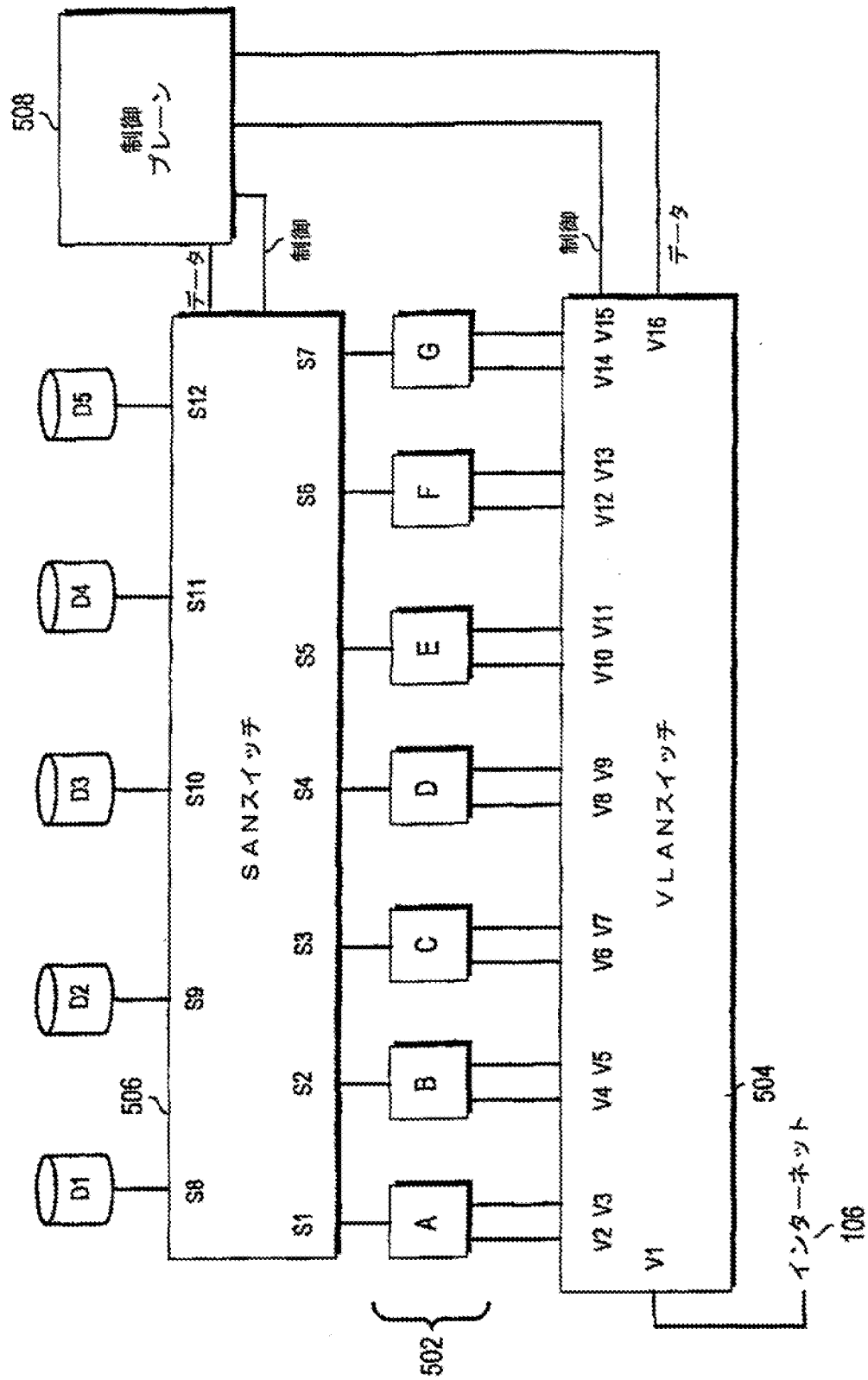
【図4C】



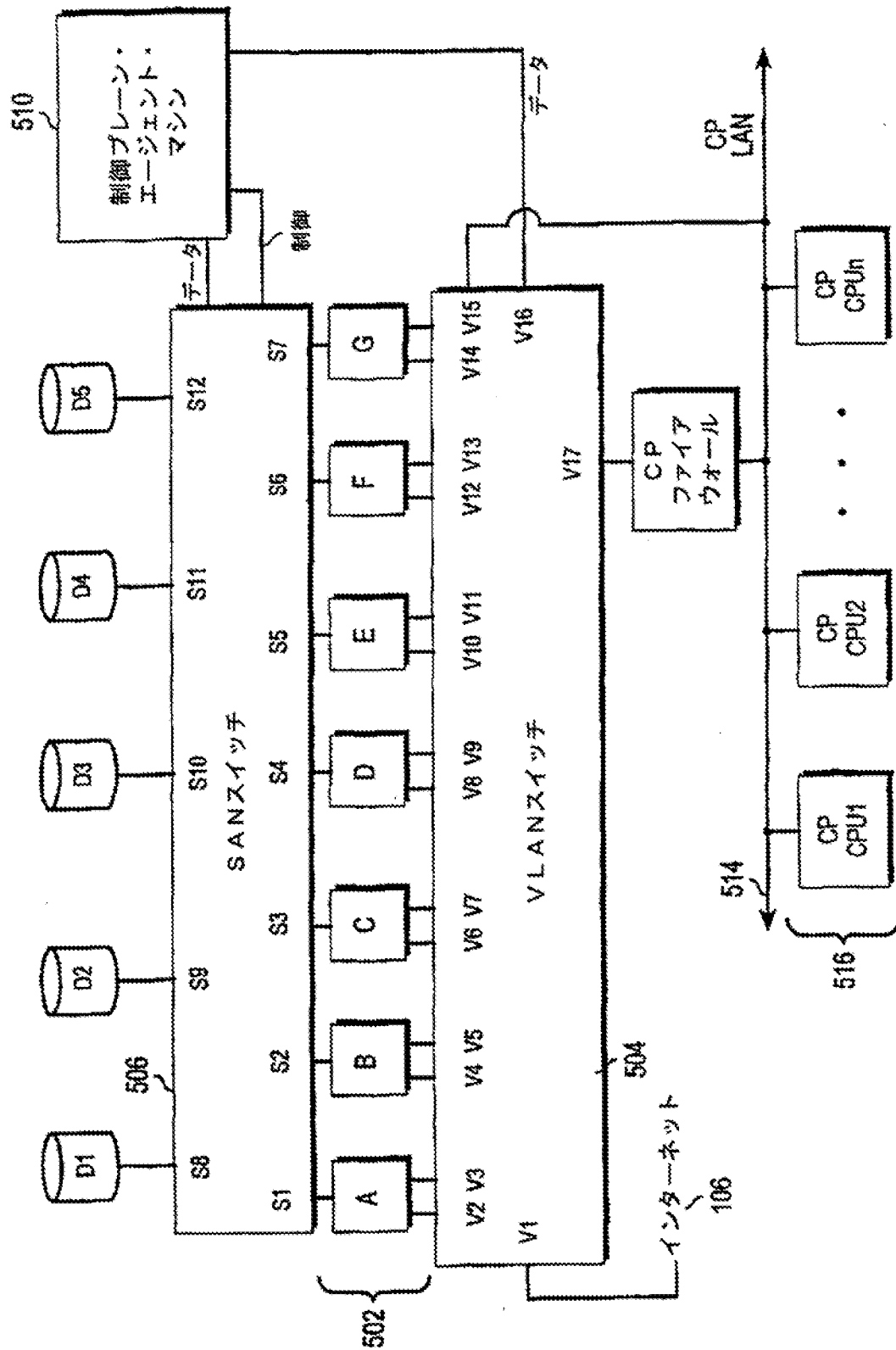
【図4D】



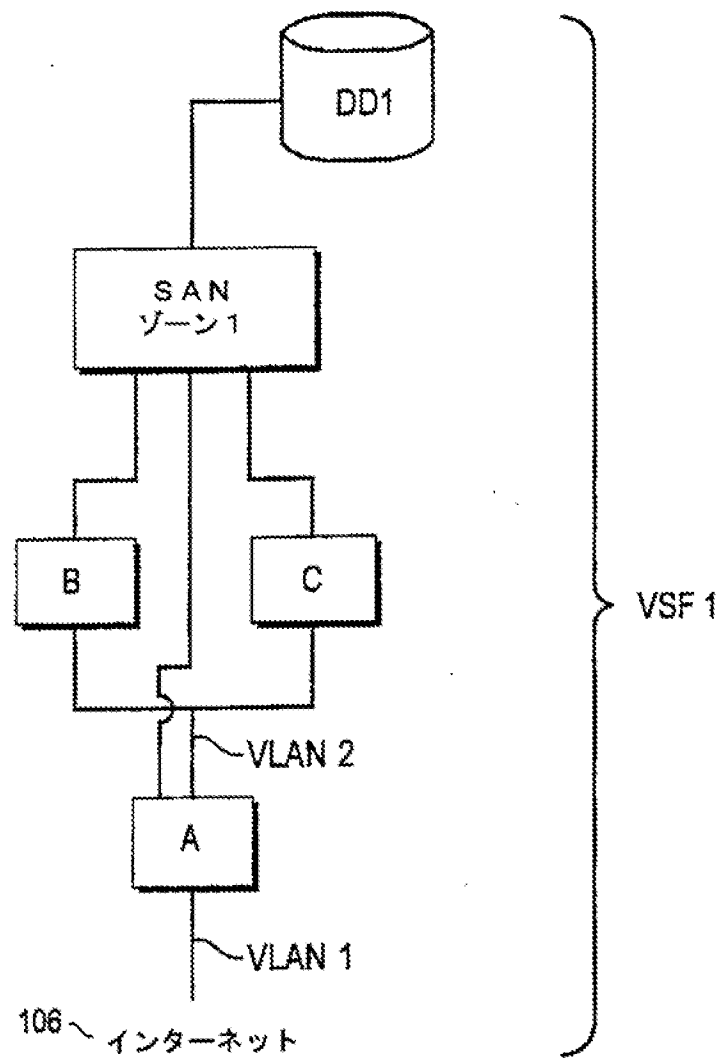
【図5A】



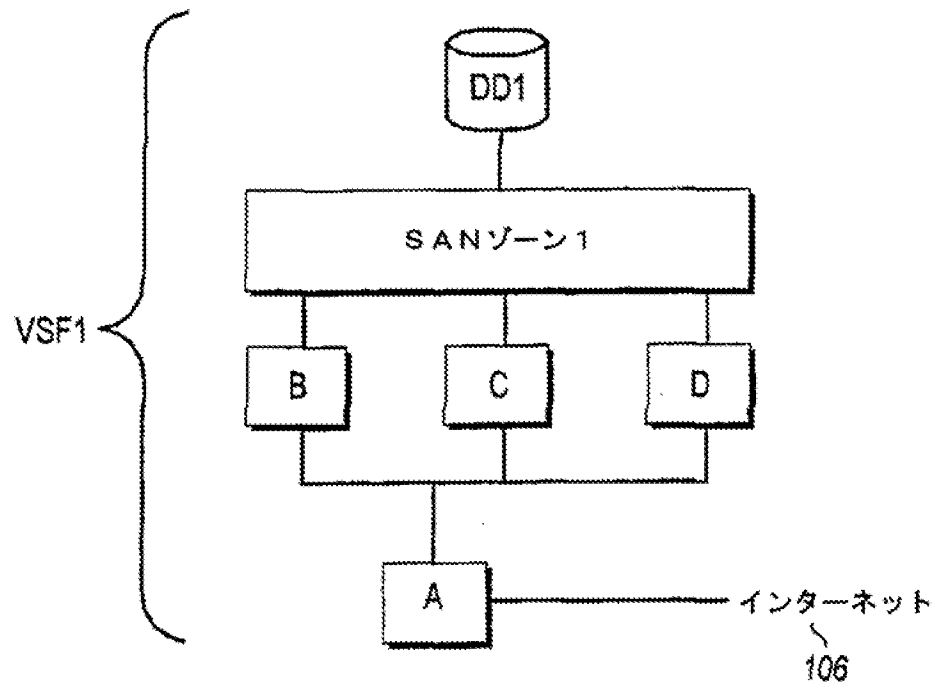
【図5B】



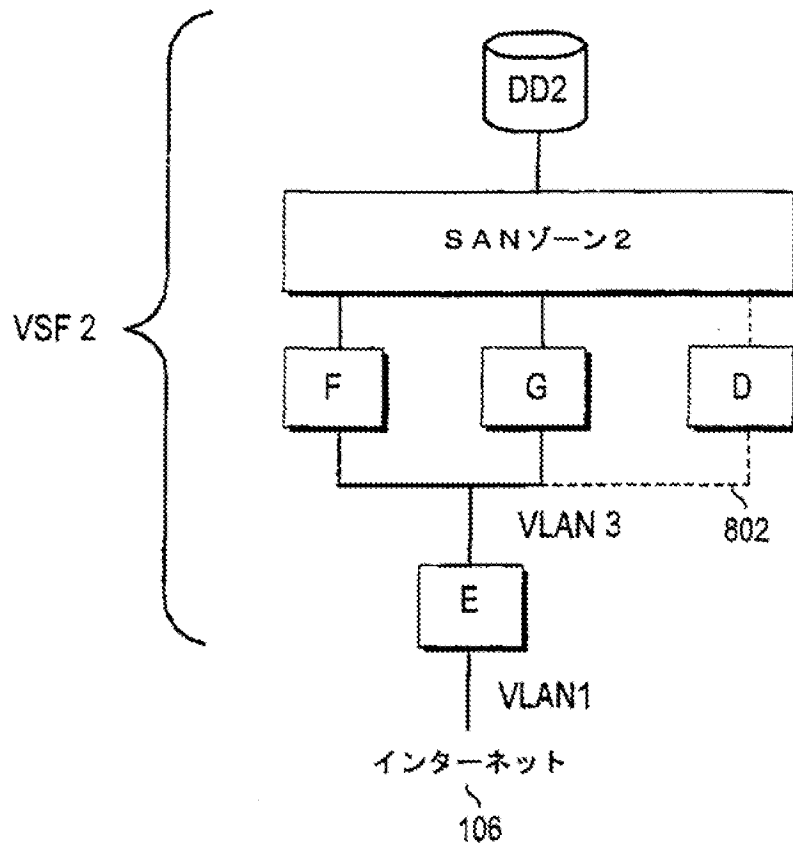
【図6】



【図7】



【図8】



【図9】

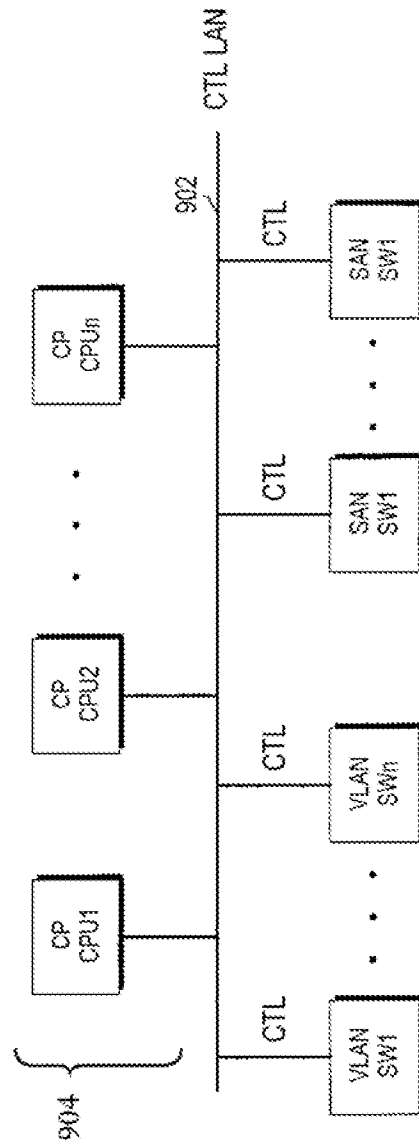
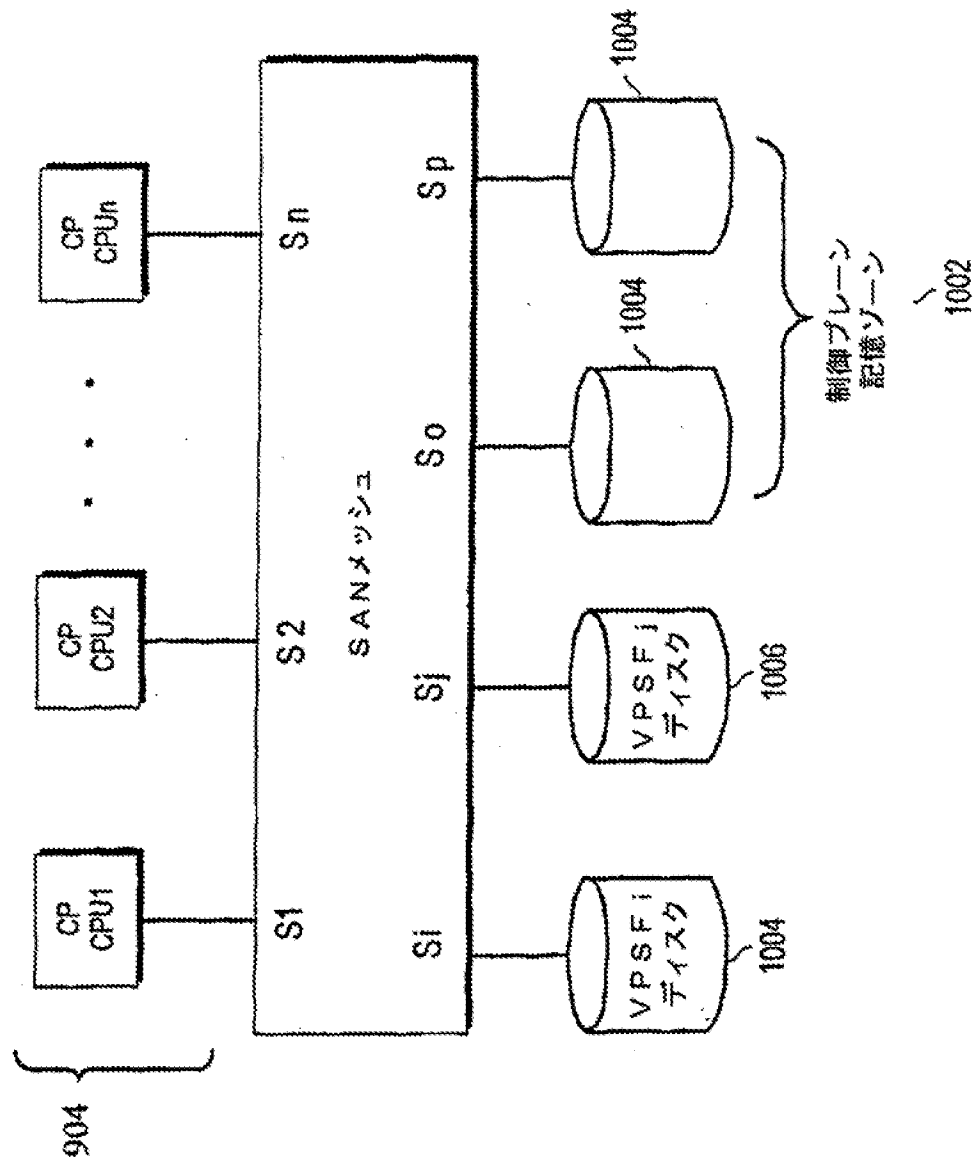
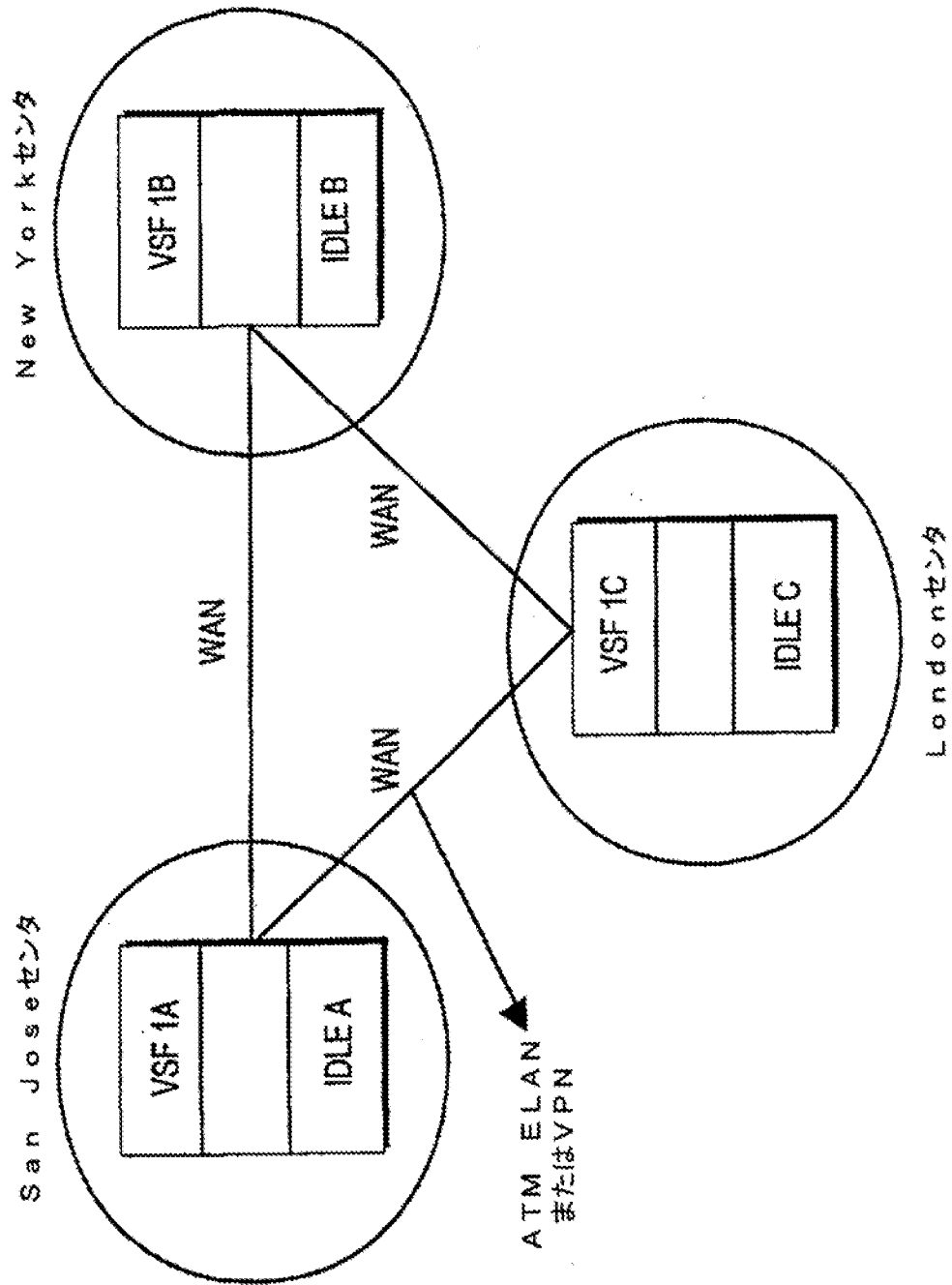


Fig. 9

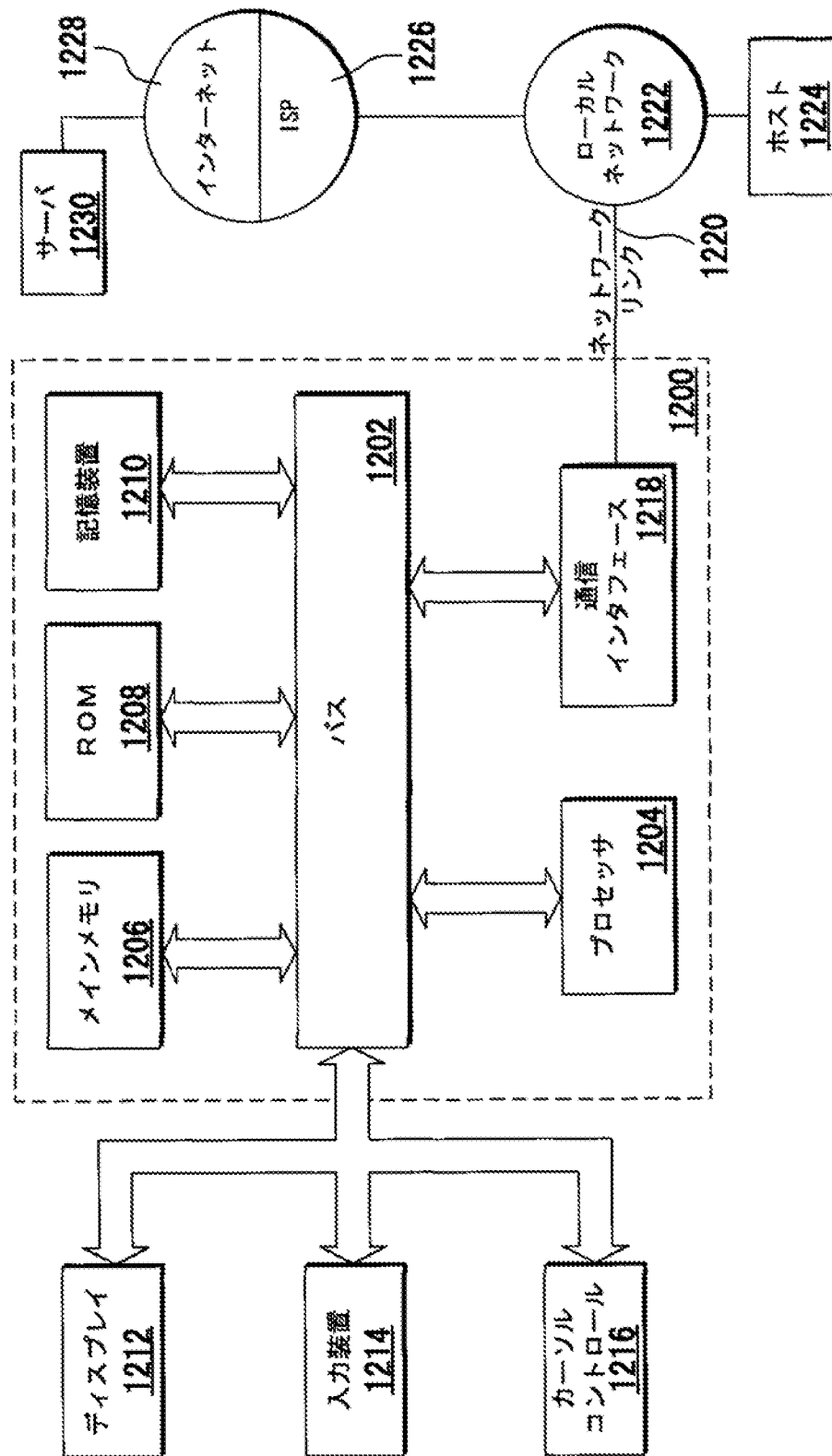
【図10】



【図11】



【図12】



【国際調査報告】

INTERNATIONAL SEARCH REPORT

International Application No.

PC/US 00/22763

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 006F9/46

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Magnetic documentation searched (classification system followed by classification symbols)
IPC 7 006F

Documentation searched other than magnetic documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (source of data base used, when practical, search terms used)

INSPEC, IBM-DB, COMPENDEX, EPD-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 750 256 A (DATA GENERAL CORP) 27 December 1996 (1996-12-27)	1,2,5, 8-14,17, 22, 24-26, 29, 32-37, 49,51
Y	abstract	3,4,15, 27,28, 36,40, 41,45
A	page 2, line 6 - line 44	3,4,6,7, 15,16, 18-21, 23,27, 28,30, 31, 38-48,50
	page 3, line 24 -page 5, line 35 -/-	

☒ Further documents are listed in the continuation of box C.☒ Patent family members are listed in annex.

* Special categories of cited documents:

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claims) or which is cited to establish the publication date of another claim or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

- "T" other document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to describe an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
- "G" document members of the same patent family

Date of the actual completion of the international search

6 March 2001

Date of mailing of the international search report

19/03/2001

Name and mailing address of the ISA

European Patent Office, P.O. Box 5018 Patentamt 2
Tel. - 2280 1111 Hgweg
Tel. (+31-70) 340-2040, Te. 31 051 800 01
Fax: (+31-70) 340-3016

Authorized officer

Ecolivet, S.

INTERNATIONAL SEARCH REPORT

International Application No.

PC1/JS 00/22783

C (Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication where appropriate of the relevant passages	Relevant to claim 14
X	NADEK VINGRALEK, YURI BREITBART, GERHARD WEIKUM: "SNOWBALL: Scalable Storage on Networks of Workstations with Balanced Load" DISTRIBUTED AND PARALLEL DATABASES, vol. 6, no. 2, April 1998 (1998-04), pages 117-156, XP002162201 Dordrecht, The Netherlands	1,2,4,5, 8,10-12, 15-17, 19,25, 26, 28-30, 33-35, 38-40, 45,46,51
A	abstract page 117, line 1 -page 119, line 30 page 120, line 1 - line 7 page 120, line 41 -page 121, line 4 page 122, line 21 -page 125, line 1 page 129, line 31 - line 39 page 130, line 27 - line 37 page 132, line 22 -page 133, line 5; figure 4 page 136, line 8 -page 137, line 7	3,6,7,9, 13,14, 18, 20-24, 27,31, 32,36, 37, 41-44, 47-50
X	US 5 574 914 A (HANCOCK PETER J ET AL) 12 November 1996 (1996-11-12) abstract; figures 3,6,7 column 2, line 8 - line 32 column 3, line 37 -column 5, line 21 column 6, line 57 - line 64	1-5,10, 23, 25-27, 29,30, 42,48,51
X	EP 0 262 750 A (THINKING MACHINES CORP) 6 April 1988 (1988-04-06)	49
Y	abstract; figures 1,2,4,5,7 column 8, line 9 - line 55 column 10, line 7 - line 12 column 10, line 23 - line 35 column 12, line 44 - line 49 column 13, line 7 - line 12 column 14, line 13 - line 20 column 14, line 48 -column 15, line 31 column 16, line 41 - line 50 column 20, line 33 -column 22, line 27	3,4,27, 28,41,45

-/-

INTERNATIONAL SEARCH REPORT

International Application No.

PC1/US 00/22783

E (Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	EP 0 905 621 A (LUCENT TECHNOLOGIES INC) 31 March 1999 (1999-03-31)	15,38,40
A	column 3, line 15 - line 26 column 4, line 37 - line 45	1,25,51
A	US 5 659 786 A (GREENSTEIN PAUL GREGORY ET AL) 19 August 1997 (1997-08-19) abstract column 1, line 47 -column 2, line 16 column 5, line 66 -column 6, line 27 column 8, line 17 -column 11, line 49	1-51
A	ARMANDO FOX, STEVEN D. GRIBBLE, YATIN CHAWATHE, ERIC A. BREWER, PAUL GAUTHIER: "CLUSTER-BASED SCALABLE NETWORK SERVICES" OPERATING SYSTEMS REVIEW (SIGOPS),US,ACM HEADQUARTER, NEW YORK, vol. 31, no. 5, 1 December 1997 (1997-12-01), pages 78-91, XP000771023 page 78, right-hand column, line 6 - line 33 page 79, right-hand column, line 27 - line 39 page 81, right-hand column, line 34 - line 57	1-51

INTERNATIONAL SEARCH REPORT

Information on patent family members

Info. National Application No.

PCI/US 00/22783

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0750256 A	27-12-1996	US 5666486 A	09-09-1997
		AU 713372 B	02-12-1999
		AU 5601396 A	09-01-1997
		CA 2179473 A	24-12-1996
		JP 9171502 A	30-06-1997
US 5574914 A	12-11-1996	NONE	
EP 0262750 A	06-04-1988	CA 1293819 A	31-12-1991
		CA 1313276 A	26-01-1993
		CN 87106067 A,B	09-03-1988
		DE 3751616 D	11-01-1996
		DE 3751616 T	09-05-1996
		IN 170067 A	01-02-1992
		JP 2792649 B	03-09-1998
		JP 63145567 A	17-06-1988
		KR 9612654 B	23-09-1996
		WO 8801772 A	10-03-1988
		US 5390336 A	14-02-1995
		US 5978570 A	02-11-1999
		US 5129077 A	07-07-1992
EP 0905621 A	31-03-1999	CA 2246867 A	26-03-1999
		JP 11161617 A	18-06-1999
US 5659786 A	19-08-1997	US 5784702 A	21-07-1998
		CA 2100540 A	20-04-1994
		EP 0593874 A	27-04-1994
		JP 7295841 A	10-11-1995

フロントページの続き

(81)指定国 EP(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OA(BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG), AP(GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), EA(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW

(72)発明者 マーティン・バターソン

アメリカ合衆国 カリフォルニア州
94041 マウンテン ビュー マーシー
ストリート 1445

Fターム(参考) 5B045 BB28 BB29 DD18 GG04 HH01
HH02 JJ22 JJ26 JJ46

【要約の続き】

一バ層、データベース・サーバ層など)は、その特定層内のサーバにかかる負荷に基づいて、動的にサイズを変更することができる。記憶デバイスは、コンピューティング・グリッド要素によって想定可能な役割に関連付けられた、複数の事前に定義された論理設計図を含むことができる。初期時には、Webサーバ、アプリケーション・サーバ、データベース・サーバなどの、任意の特定の役割またはタスク専用のコンピューティング要素はない。各コンピューティング要素の役割は、複数の事前に定義され記憶された設計図の1つから取得されるものであって、設計図はそれぞれが、役割に関連付けられたコンピューティング要素にブート・イメージを定義するものである。

